LARRY GONICK
ART HUFFMAN

LES STATISTIQUES EN BD

GRÂCE À LARRY GONICK, LES SCIENCES RETROUVENT LE SOURIRE ET LES CONCEPTS DIFFICILES DEVIENNENT LIMPIDES.

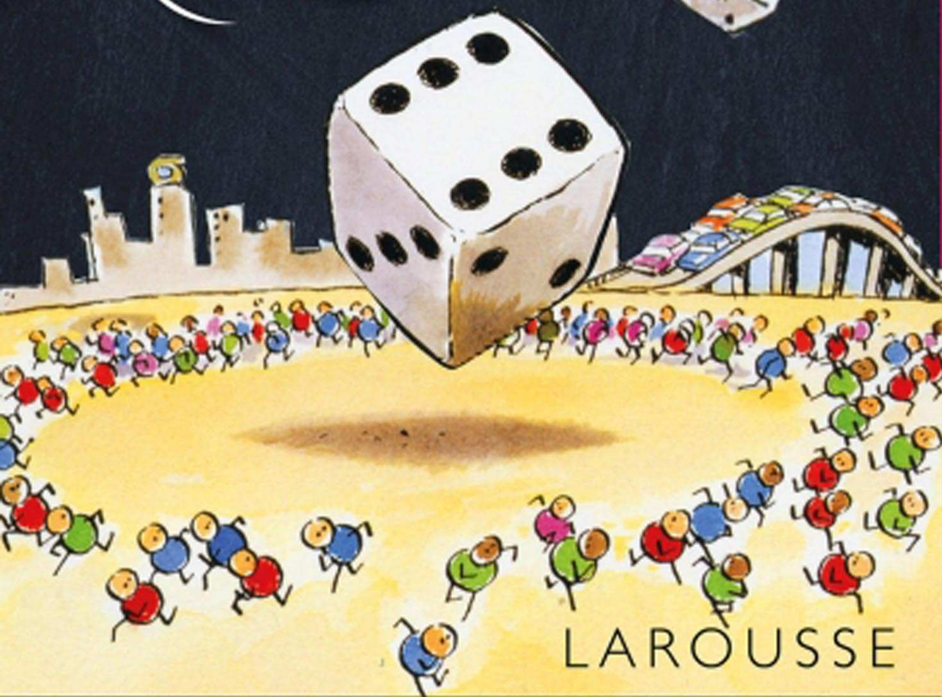












$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ ou } \sum_{i=1}^{n} \frac{x_i}{n}$$

LAROUSSE

Les Statistiques en BD

Autres ouvrages de Larry Gonick :

Aux Éditions Larousse (en français) :

La Chimie en BD
Les Maths en BD
La Génétique en BD
La Physique en BD

Et en anglais :

The Cartoon History of the Universe, Volumes 1-7
The Cartoon History of the Universe II, Volumes 8-13
The Cartoon History of the Universe III, Volumes 14-19
The Cartoon History of the United States
The Cartoon Guide to the Computer
The Cartoon Guide to the Environment (avec Alice Outwater)
The Cartoon Guide to (Non) Communication
The Cartoon Guide to Sex (avec Christine Devault)

LES STATISTIQUES EN BD





LARRY GONICK
& WOOLLCOTT SMITH

LAROUSSE

Édition originale :

Ce titre a été publié pour la première fois en 1993 sous le titre original
The Cartoon Guide to Statistics © 1993 by Larry Gonick and Woollcott Smith.
Published by arrangement with William Morrow, an imprint of HARPERCOLLINS PUBLISHERS.

Édition française :

Éditions Larousse © 2016 pour l'édition en langue française

Direction de la publication : Carine Girac-Marinier

Direction éditoriale : Claude Nimmo

Direction éditoriale adjointe : Julie Pelpel-Moulian

Édition : Antoine Caron

Réalisation et coordination éditoriale : Belle Page

Traduction : Thierry Lafay, maître de conférences à l'université Paris 1 Panthéon-Sorbonne. Il enseigne notamment les mathématiques, les statistiques et les techniques quantitatives à l'École de Management de la Sorbonne.

ISBN : 978-2-03-593291-4

Sommaire

À PROPOS DES AUTEURS

WOOLCOTT SMITH EST PROFESSEUR ÉMÉRITE DE STATISTIQUES À L'UNIVERSITÉ DE TEMPLE. TITULAIRE D'UNE LICENCE ET D'UN MASTÈRE DE L'UNIVERSITÉ DU MICHIGAN ET D'UNE THÈSE DE L'UNIVERSITÉ JOHNS- HOPKINS, IL EST AUTEUR ET COAUTEUR DE PLUS D'UNE QUARANTAINE DE PUBLICATIONS DANS DIFFÉRENTS DOMAINES TELS QUE LES MARÉES NOIRES, LA THÉORIE STATISTIQUE ET LES STATISTIQUES ENVIRONNEMENTALES. IL A CONSEILLÉ DE NOMBREUX PROGRAMMES SCIENTIFIQUES NATIONAUX. IL ENTRETIENT DES DÉBATS ANIMÉS ET FAIT DU KAYAK AVEC SA FEMME LEAH ET SES DEUX ENFANTS MAINTENANT ADULTES KESTON ET AMELIA.





LARRY GONICK EST L'AUTEUR ET LE COAUTEUR DE GUIDES ILLUSTRÉS ET DE BANDES DESSINÉES AYANT REÇU LE PRIX HARVEY. IL A ÉCRIT ET DESSINÉ LES CLASSIQUES DE LA SCIENCE DANS *DISCOVER MAGAZINE* ET A CRÉÉ LA BANDE DESSINÉE *KOKOPELLI & COMPANY* POUR LE MAGAZINE *MUSE*. IL EST L'UNIQUE DESSINATEUR À AVOIR CONTRIBUÉ AVEC DES DESSINS COMIQUES ORIGINAUX À LA PRESTIGIEUSE REVUE *SCIENCE*. APRÈS AVOIR ABANDONNÉ LES MATHÉMATIQUES À LA HARVARD SCHOOL, IL S'EST INSTALLÉ SUR LA CÔTE PACIFIQUE AVEC SA FEMME LISA. LEURS DEUX FILLES SOPHIE ET ANNA ONT DÉSORMAIS QUITTÉ LE COCON FAMILIAL.

Remerciements

Nous serons éternellement reconnaissants envers notre éditeur original Carol Cohen de Happercollins qui nous a suggéré ce projet et envers notre agente littéraire Vicky Bijur qui nous a permis de le mener à bien grâce aux coauteurs. Un coup de chapeau particulier aux éditeurs et assistants avec lesquels nous avons travaillé depuis : Erica Spaberg, Stephanie Meyer, Peter Hubbard, Cole Hager et Nick Amphlett.

Les commentaires de William Fairley et Leah Smith ont contribué à l'amélioration des versions préliminaires.

Donna Okino a fourni une assistance et des conseils inestimables pour la production et la préparation des pages dessinées. Elle dit que créer un guide illustré est plus difficile que de courir un marathon et elle sait de quoi elle parle : elle a fait les deux.

Des remerciements particuliers aux créateurs de Fontographer, le logiciel magique qui nous a permis de créer plusieurs polices de caractères qui simulent l'écriture manuelle.

Et comme l'éducation n'est pas à sens unique, un coup de chapeau à la patience des étudiants de l'université de Temple. Le futur leur appartient.

Note du traducteur : dans ce livre, nous avons conservé les données du manuscrit original bien que les unités soient américaines. En effet, les statistiques travaillent sur les chiffres et peu importe l'unité choisie. Il ne nous semblait donc pas nécessaire de convertir tous ces jeux de données, choisis par les auteurs pour leur intérêt statistique et illustrés par des dessins originaux. Nous avons toutefois donné le ratio pour convertir ces chiffres en unités françaises, les conclusions étant aussi agrémentées de leur conversion.



Chapitre I
Qu'est-ce que la statistique?

ON SE DÉBROUILLE DANS LA VIE EN FAISANT DES CHOIX FONDÉS SUR UNE **INFORMATION INCOMPLÈTE...**

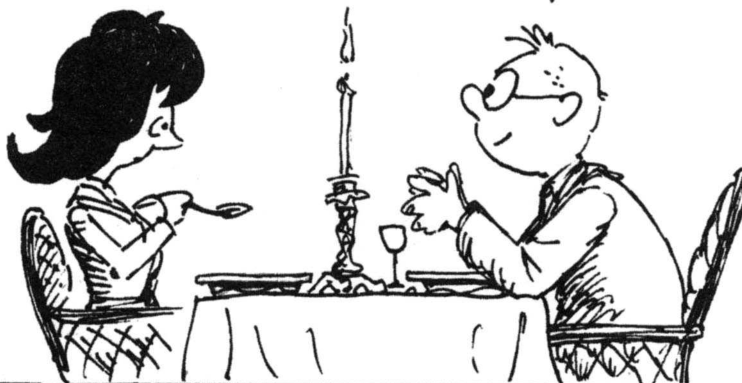


La plupart d'entre nous vivent sans problème avec ce niveau d'incertitude.





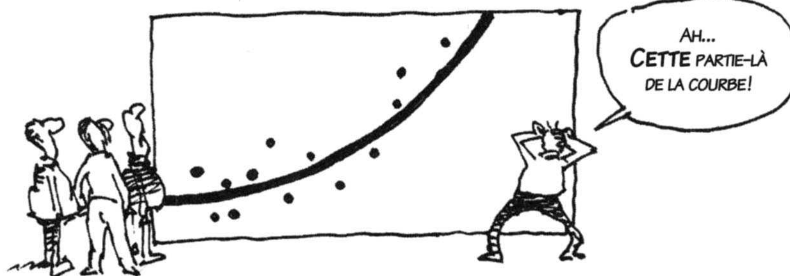
Ce qui rend les statistiques si uniques, c'est la possibilité de **quantifier** l'incertitude, de façon à la rendre plus précise. Les statisticiens peuvent alors faire avec assurance des **propositions catégoriques** sur leur niveau d'incertitude !



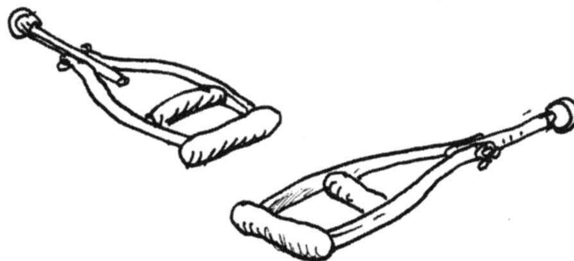
CELA NE SERT PAS UNIQUEMENT À COMMANDER UNE SOUPE ! LA STATISTIQUE CONCERNE AUSSI DES QUESTIONS **DE VIE OU DE MORT...**



PAR EXEMPLE, EN 1986, LA NAVETTE SPATIALE **CHALLENGER** A EXPLOSÉ, TUANT SEPT ASTRONAUTES. LA DÉCISION DE LANCER LA NAVETTE ALORS QU'IL FAISAIT - 1 °C A ÉTÉ PRISE SANS QU'AUCUNE ANALYSE SIMPLE AIT ÉTÉ FAITE SUR LES PERFORMANCES DE LA NAVETTE À BASSE TEMPÉRATURE.



LE **VACCIN DE SALK** CONTRE LA POLIOMYÉLITE FOURNIT UN EXEMPLE PLUS POSITIF. EN 1954, DES ESSAIS CLINIQUES DU VACCIN FURENT TESTÉS SUR PLUS DE 400 000 ENFANTS, AVEC UN CONTRÔLE STRICT POUR ÉLIMINER DES RÉSULTATS BIAISÉS. LES ANALYSES STATISTIQUES DES RÉSULTATS ONT ÉTÉ CONCLUANTES. ELLES ONT PERMIS D'ÉTABLIR L'EFFICACITÉ DU VACCIN ET, AUJOURD'HUI, LA POLIOMYÉLITE EST PRESQUE ÉRADIQUÉE.



POUR ACCOMPLIR LEURS EXPLOITS DE DÉTECTIVE DU MONDE RÉEL, LES STATISTICIENS UTILISENT TROIS DISCIPLINES LIÉES :

L'analyse des données

LA COLLECTE, LA PRÉSENTATION ET LE RÉSUMÉ DES DONNÉES.

Les probabilités

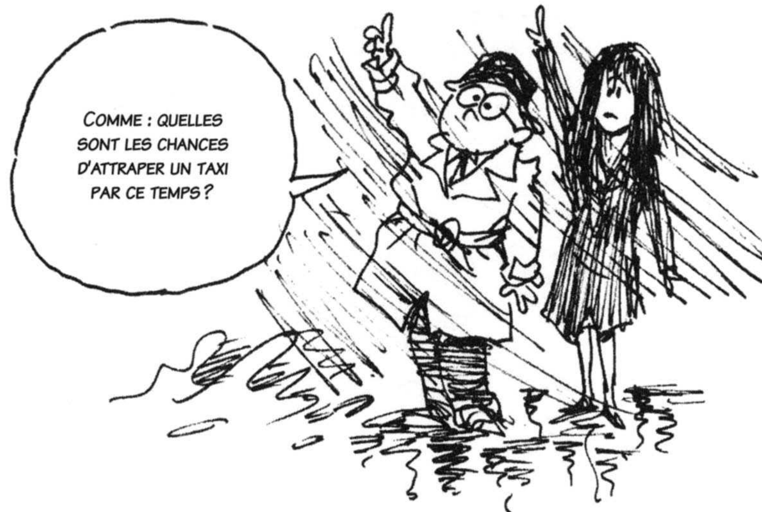
LES LOIS DU HASARD, QUE CE SOIT DANS OU EN DEHORS D'UN CASINO.

L'inférence statistique

LA SCIENCE QUI CONSISTE À ÉLABORER DES CONCLUSIONS STATISTIQUES À PARTIR DE DONNÉES SPÉCIFIQUES EN UTILISANT LES CONNAISSANCES SUR LES PROBABILITÉS.



DANS CET OUVRAGE, NOUS ALLONS ABORDER CES TROIS DISCIPLINES EN LES APPLIQUANT À DE NOMBREUX TYPES DE SITUATIONS OÙ LES STATISTIQUES JOUENT UN RÔLE CRUCIAL DANS LE MONDE MODERNE.



DANS LE CHAPITRE 2, NOUS EXAMINERONS UN ENSEMBLE SIMPLE DE DONNÉES : LES POIDS D'UN GROUPE D'ÉTUDIANTS DE LICENCE.



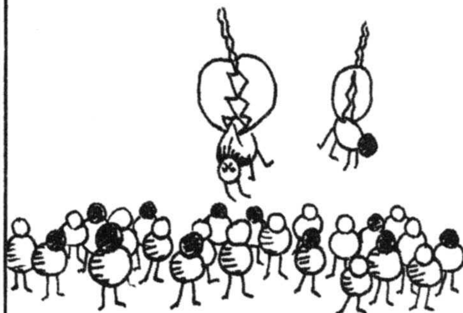
DANS LE CHAPITRE 3, NOUS ÉTUDIERONS LES LOIS DE PROBABILITÉS LÀ OÙ ELLES SONT APPARUES, LES MAISONS DE JEUX.



LES CHAPITRES 4 ET 5 MONTRERONT COMMENT DÉCRIRE LE MONDE GRÂCE AUX **MODÈLES DE PROBABILITÉS**, ET AU CONCEPT DE **VARIABLE ALÉATOIRE.**



LE CHAPITRE 6 INTRODUIRA L'UNE DES PREMIÈRES PROCÉDURES DU BON STATISTICIEN : CONSTRUIRE UN **ÉCHANTILLON** À PARTIR D'UNE GRANDE POPULATION.



DANS LE CHAPITRE 7 ET LES SUIVANTS, NOUS DÉCRIRONS COMMENT FAIRE DES INFÉRENCES STATISTIQUES DANS DES SITUATIONS RÉELLES AUSSI COURANTES QUE LES **PRÉVISIONS ÉLECTORALES**, LE **CONTRÔLE DE QUALITÉ INDUSTRIELLE**, LES **TESTS MÉDICAUX**, LE **SUIVI ENVIRONNEMENTAL**, LES **DISCRIMINATIONS** ET LE **DROIT**.



ENFIN, LORSQUE NOUS PARLONS DE STATISTIQUES IL EST DIFFICILE D'ÉVITER DE MENTIONNER LA **MÉFIANCE** TRÈS RÉPANDUE À L'ÉGARD DES STATISTIQUES AUJOURD'HUI. ON A TOUS ENTENDU DIRE QUE LES STATISTIQUES SONT LA « FORME LA PLUS ÉLABORÉE DU MENSONGE », ET IL EST PRESQUE IMPOSSIBLE DE TROUVER DE BONNES ANALYSES STATISTIQUES DANS LA VIE DE TOUS LES JOURS. ALORS QUE FAIRE ?




NOTRE HUMBLE OPINION EST QU'EN **APPRENDRE UN PEU PLUS SUR LE SUJET** N'EST SÛREMENT PAS UNE MAUVAISE IDÉE... ET C'EST POUR CELA QUE NOUS AVONS ÉCRIT CE LIVRE !



DANS CE QUI SUIT, NOUS ESSAIERONS DE PRÉSENTER LES ÉLÉMENTS STATISTIQUES DE LA FAÇON LA PLUS VISUELLE ET INTUITIVE POSSIBLE. TOUT CE DONT VOUS AVEZ BESOIN EST D'UN SOUPÇON DE **PATIENCE,** D'UN PEU DE **RÉFLEXION** ET D'UNE CERTAINE TOLÉRANCE À L'**ALGÈBRE** OU, SI CE N'EST PAS LE CAS, CE SERAIT PEUT-ÊTRE UN BON **PRÉALABLE À CE COURS** !

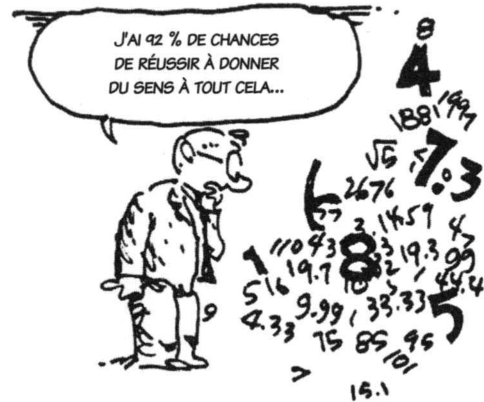




Chapitre 2
Statistiques descriptives



LES **DONNÉES** CONSTITUENT LA MATIÈRE BRUTE DES STATISTICIENS. NOUS UTILISONS DES NOMBRES POUR TRANSCRIRE LA RÉALITÉ. TOUS LES PROBLÈMES STATISTIQUES IMPLIQUENT UNE COLLECTE, UNE DESCRIPTION ET UNE ANALYSE DES DONNÉES **OU UNE RÉFLEXION** SUR CETTE COLLECTE, SUR CETTE DESCRIPTION ET SUR L'ANALYSE DES DONNÉES.



CE CHAPITRE SE CONCENTRE SUR LA PARTIE **DESCRIPTIVE** DES DONNÉES. COMMENT REPRÉSENTER LES DONNÉES DE FAÇON PRATIQUE ET UTILE ? COMMENT SOULIGNER DES TENDANCES À PARTIR D'UNE ACCUMULATION DE DONNÉES BRUTES ? COMMENT SYNTHÉTISER ET RÉSUMER LA FORME BASIQUE DE CES DONNÉES ?



EH BIEN, POUR DÉCRIRE LES DONNÉES, NOUS AVONS BESOIN AVANT TOUT DE DISPOSER DE DONNÉES... ALORS, COMMENÇONS LEUR COLLECTE !



Voici des données réellement collectées lors d'une expérience en classe. Elles concernent le poids en livres de 92 étudiants de l'université de Penn State :



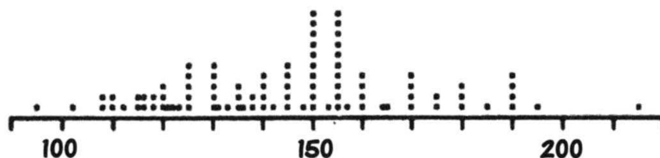
Poids des 57 étudiants

140 145 160 190 155 165 150 190 195 138 160 155 153 145 170 175 175 170 180
135 170 157 130 185 190 155 170 155 215 150 145 155 155 150 155 150 180 160
135 160 130 155 150 148 155 150 140 180 190 145 150 164 140 142 136 123 155

Poids des 35 étudiantes

140 120 130 138 121 125 116 145 150 112 125 130 120 130 131 120 118 125 135
125 118 122 115 102 115 150 110 116 108 95 125 133 110 150 108

Revenons-en aux choses sérieuses et traçons un graphique de points : pour chaque étudiant on associe un point correspondant à son poids en livres. Et on empile les points lorsque plusieurs étudiants ont donné la même réponse.



Poids en livres (1 livre = 0,453 kg)



Vous pouvez noter ici un **problème** : les amas de points sur 150 et 155 livres. Les étudiants ont eu tendance à communiquer leur poids en **arrondissant à 5 livres** près. Dans la réalité, ce type d'arrondi peut nuire à l'analyse de tendance sur les données... Mais pour le moment, nous allons travailler avec celles-ci.

ON PEUT RÉSUMER LES DONNÉES EN UTILISANT UN **TABLEAU DES EFFECTIFS.** ON REGROUPE ALORS LES DONNÉES PAR TRANCHES (NOMMÉES « CLASSES ») ET ON COMPTE LE NOMBRE DE POIDS D'ÉTUDIANTS DANS CHAQUE INTERVALLE. L'**EFFECTIF** CORRESPOND AU NOMBRE D'ÉTUDIANTS POUR CHAQUE CLASSE. LA **FRÉQUENCE** EST ALORS LA PROPORTION DES POIDS DANS CHAQUE INTERVALLE. IL S'AGIT DES EFFECTIFS DIVISÉS PAR LE NOMBRE TOTAL D'ÉTUDIANTS.

CLASSE	CENTRE DE CLASSE	EFFECTIF	FRÉQUENCE
87,5–102,4	95	2	0,022
102,5–117,4	110	9	0,098
117,5–132,4	125	19	0,206
132,5–147,4	140	17	0,185
147,5–162,4	155	27	0,293
162,5–177,4	170	8	0,087
177,5–192,4	185	8	0,087
192,5–207,4	200	1	0,011
207,5–222,4	215	1	0,011
TOTAL		92	1,000

NOTE : NOUS AVONS FIXÉ, POUR LES INTERVALLES, DES BORNES ÉLOIGNÉES DES PROBLÈMES D'ARRONDIS À 5 LIVRES PRÈS. CELA PERMET DE LIMITER LES BIAIS VENANT DES DÉCLARATIONS DES ÉTUDIANTS.

DIRECTIVES POUR CRÉER DES CLASSES :

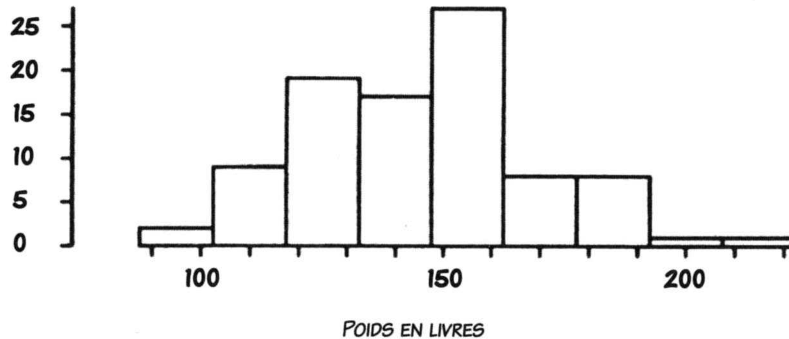
1) UTILISER DES CLASSES DE MÊME AMPLITUDE AVEC DES CENTRES AUX VALEURS ARRONDIES APPROPRIÉES.

2) POUR UN PETIT NOMBRE DE DONNÉES, UTILISER UN PETIT NOMBRE DE CLASSES.

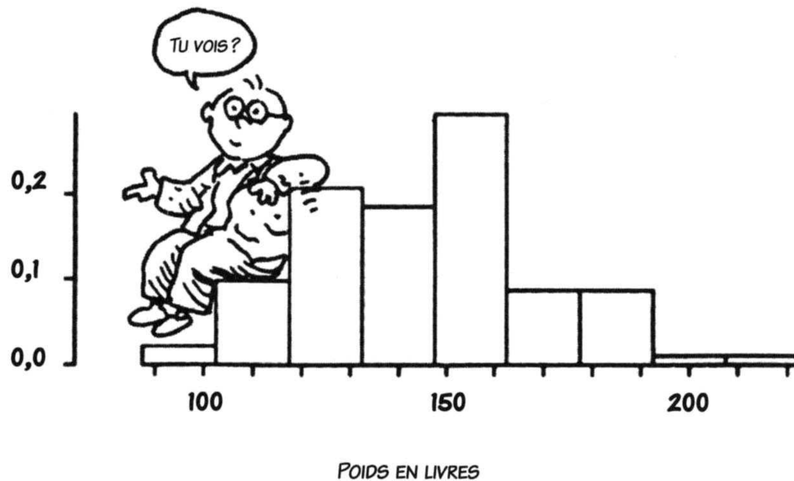
3) POUR UN GRAND NOMBRE DE DONNÉES, UTILISER PLUS DE CLASSES.



DANS LE TABLEAU DES EFFECTIFS, NOUS MONTRONS COMBIEN DE DONNÉES SE TROUVENT « AUTOUR » DE CHAQUE CENTRE DE CLASSE. ON PEUT AUSSI FAIRE UN DESSIN POUR REPRÉSENTER CE TYPE D'INFORMATION. LE DIAGRAMME EN BARRE ASSOCIÉ EST APPELÉ **HISTOGRAMME.** CHAQUE BARRE REPRÉSENTE UNE CLASSE ET EST CENTRÉE SUR LE CENTRE DE CLASSE. LA HAUTEUR DU RECTANGLE EST LE NOMBRE DE DONNÉES DANS LA CLASSE.



ON PEUT AUSSI DESSINER UN **HISTOGRAMME DES FRÉQUENCES,** EN REPORTANT LA FRÉQUENCE POUR CHAQUE CLASSE DE POIDS. LE GRAPHE EST EXACTEMENT LE MÊME, SEULE L'ÉCHELLE DE L'AXE VERTICAL CHANGE.



LE STATISTICIEN **JOHN TUKEY** (1915-2000) A INVENTÉ UNE MÉTHODE RAPIDE POUR RÉSUMER DES DONNÉES TOUT EN CONSERVANT LES INFORMATIONS BRUTES. IL S'AGIT DU DIAGRAMME **BRANCHE ET FEUILLE.**



POUR NOS DONNÉES DE POIDS, LA BRANCHE EST UNE COLONNE DE NOMBRES, CORRESPONDANT AUX POIDS EN LIVRES RANGÉS PAR DIZAINES (C'EST-À-DIRE EN OMETTANT LE DERNIER CHIFFRE).

9
10
11
12
13
14
15
16
17
18
19
20
21



POUR LE POIDS DES **ÉTUDIANTES** DE LA PAGE 9, ON AJOUTE LE DERNIER CHIFFRE DE CHAQUE POIDS DANS LA LIGNE APPROPRIÉE.

BRANCHE : FEUILLES

9 : 5
10 : 288
11 : 628855060
12 : 0155005525
13 : 0800153
14 : 05
15 : 000
16 :
17 :
18 :
19 :
20 :
21 :



ENSUITE, ON PEUT RANGER LES « FEUILLES » PAR ORDRE CROISSANT.

BRANCHE : FEUILLES

9 : 5
10 : 288
11 : 002556688
12 : 0001255555
13 : 0001358
14 : 05
15 : 000
16 :
17 :
18 :
19 :
20 :
21 :

DE LA MÊME MANIÈRE, ON PEUT FAIRE LE DIAGRAMME BRANCHE ET FEUILLE DES POIDS DES **ÉTUDIANTS HOMMES.**

9 :
10 :
11 :
12 : 13
13 : 005568
14 : 000255558
15 : 0000000355555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5



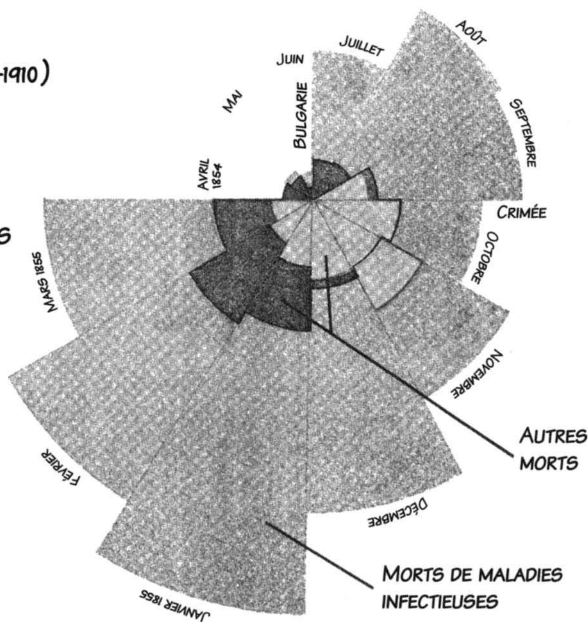
TOUS CES 0 ET CES 5 INDIQUENT CLAIREMENT LE BIAIS DÛ AUX INDICATIONS ARRONDIES DES ÉTUDIANTS.

Un bon affichage graphique doit être à la fois artistique et scientifique.



Et parfois **politique !**

L'infirmière dévouée **Florence Nightingale** (1820-1910) a compilé les **statistiques de mortalité** dans les hôpitaux militaires britanniques. Elle a produit des histogrammes troublants comme celui-ci : le rayon augmente avec le nombre de soldats britanniques morts lors de la guerre de Crimée (que ce soit dans les hôpitaux ou sur le champ de bataille). La plupart de ces morts étaient, et de loin, dues à des maladies infectieuses.

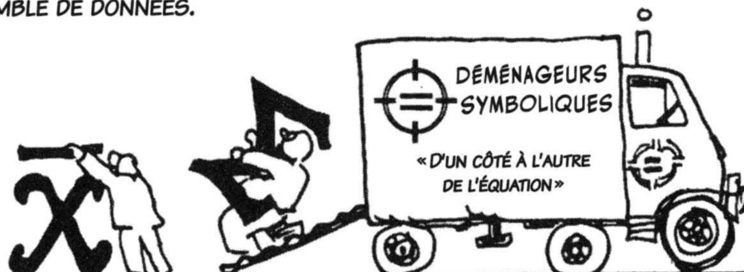


Ses efforts statistiques ont directement permis d'améliorer les conditions à l'hôpital. Ce qui a entraîné une réduction du taux de décès.

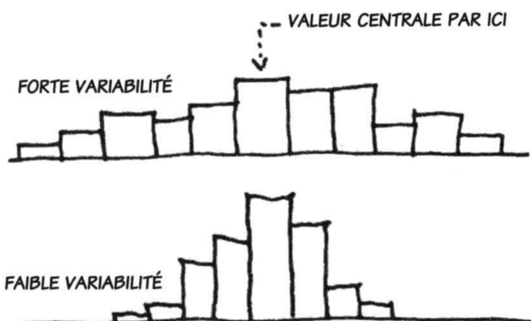


RÉSUMÉ STATISTIQUE NUMÉRIQUE

MAINTENANT, PASSONS DES IMAGES AUX FORMULES. NOTRE OBJECTIF EST D'OBTENIR DES MESURES SIMPLES DES CARACTÉRISTIQUES RUDIMENTAIRES D'UN ENSEMBLE DE DONNÉES.



TOUS LES TYPES DE MESURES ONT DEUX DIMENSIONS IMPORTANTES : LA **VALEUR** OU **TENDANCE CENTRALE,** ET LA **VARIABILITÉ** AUTOUR DE CETTE VALEUR. CETTE IDÉE APPARAÎT SUR LES DEUX HISTOGRAMMES HYPOTHÉTIQUES REPRÉSENTÉS.



ON PEUT FAIRE BEAUCOUP DE CHOSES AVEC UN PETIT PEU DE NOTATION. SUPPOSONS QUE L'ON AIT UNE SÉRIE D'OBSERVATIONS... *n* POUR ÊTRE PRÉCIS... ALORS ON PEUT NOTER :

$$x_1, x_2, x_3 \dots x_n$$

LES DIFFÉRENTES VALEURS OBSERVÉES. AINSI, *n* EST LE NOMBRE TOTAL D'OBSERVATIONS, ET x_4 (PAR EXEMPLE) EST LA VALEUR DE LA QUATRIÈME OBSERVATION.



UNE **MATRICE** EST UN TABLEAU DE DONNÉES :

OBSERVATION	1	2	3	4 ... n
VALEUR OBSERVÉE	x_1	x_2	x_3	$x_4 \dots x_n$

SUR UN PETIT ENSEMBLE DE $n = 5$ DONNÉES, ON PEUT TOUT FAIRE À LA MAIN. PAR EXEMPLE, ON DEMANDE À 5 PERSONNES COMBIEN D'HEURES PAR SEMAINE ILS PASSENT DEVANT LA **TÉLÉVISION...** ET VOICI LA MATRICE DES RÉSULTATS :

OBSERVATION	1	2	3	4	5
VALEUR OBSERVÉE	5	7	3	38	7

ALORS $x_1 = 5$, $x_2 = 7$, $x_3 = 3$, $x_4 = 38$, $x_5 = 7$

QUEL EST LE « CENTRE » DE CES DONNÉES ? EN FAIT, IL EXISTE DIFFÉRENTES RÉPONSES À CETTE QUESTION. NOUS ALLONS EN EXAMINER DEUX.



LA MOYENNE

LA **MOYENNE** EST REPRÉSENTÉE PAR $\bar{x}$. ELLE EST OBTENUE EN AJOUTANT TOUTES LES DONNÉES OBSERVÉES ET EN DIVISANT PAR LE NOMBRE D'OBSERVATIONS.

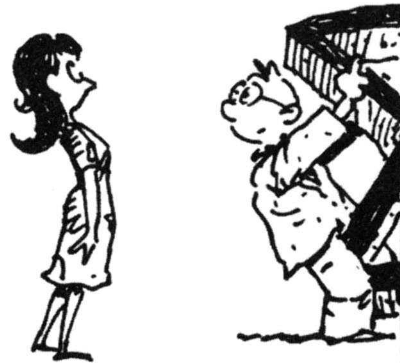
$$\bar{x} = \frac{\text{SOMME DES DONNÉES}}{n}$$

$$= \frac{x_1 + x_2 + \dots + x_n}{n}$$

POUR NOTRE EXEMPLE,

$$\bar{x} = \frac{5 + 7 + 3 + 38 + 7}{5} = \frac{60}{5}$$

= **12 HEURES**



IL EXISTE UNE NOTATION MATHÉMATIQUE PARTICULIÈRE POUR LA SOMME $x_1 + x_2 + ... + x_n$. ON UTILISE DANS CE CAS LA LETTRE GRECQUE **SIGMA.**

$$\Sigma$$

AU LIEU DE $x_1 + x_2 + ... + x_n$, ON ÉCRIT :

$$\sum_{i=1}^{n} x_i$$

ET ON LIT « SOMME DE x_i POUR i ÉGAL 1 À n ».





AINSI... LA **MOYENNE** D'UN ENSEMBLE DE DONNÉES x_i EST :

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{ou} \quad \sum_{i=1}^{n} \frac{x_i}{n}$$

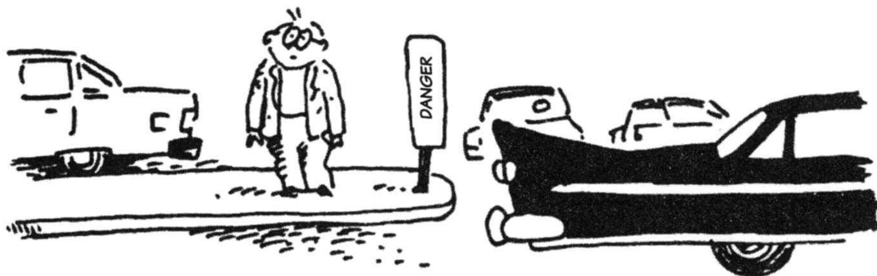
DANS LE CAS DES 92 ÉTUDIANTS DE PENN STATE, LE POIDS MOYEN EST :

$$\sum_{i=1}^{92} = \frac{x_i}{92} = \frac{13\,354}{92}$$



145,15 LIVRES, SOIT ENVIRON 65,84 kg.

LA **MÉDIANE** EST UNE AUTRE MESURE DE TENDANCE CENTRALE : IL S'AGIT DU « MILIEU » QUI PARTAGE EN DEUX L'ENSEMBLE DES DONNÉES, TOUT COMME LA MÉDIANE D'UN TRIANGLE OU LE TERRE-PLEIN CENTRAL D'UNE AVENUE.



POUR TROUVER LA MÉDIANE D'UN ENSEMBLE DE DONNÉES, IL FAUT TRIER PAR ORDRE CROISSANT LES DONNÉES. LA MÉDIANE EST ALORS LA VALEUR CENTRALE.

3 5 7 7 38

LA MÉDIANE (→ 7)

SI LE NOMBRE D'OBSERVATIONS EST PAIR, IL N'Y A PAS VRAIMENT DE VALEUR CENTRALE. ON PREND ALORS LA MOYENNE DES DEUX VALEURS AUTOUR DU MILIEU...
AINSI SI LES DONNÉES SONT :

3 5 7 7

ESPACE CENTRAL

ON FAIT LA MOYENNE ENTRE 5 ET 7 ET ON TROUVE $\frac{5+7}{2} = 6$

CELA NOUS DONNE LA RÈGLE GÉNÉRALE DE CALCUL : ON TRIE PAR ORDRE CROISSANT LES DONNÉES.

SI LE NOMBRE D'OBSERVATIONS EST **IMPAIR**, LA MÉDIANE EST LA VALEUR CENTRALE.

SI LE NOMBRE D'OBSERVATIONS EST **PAIR**, LA MÉDIANE EST LA MOYENNE DES DEUX DONNÉES CENTRALES SITUÉES AUTOUR DU MILIEU.



Pour les $n = 92$ poids d'étudiants, on peut calculer la médiane à partir du diagramme branche et feuille triée. Il suffit de compter jusqu'à la 46e observation. La médiane est alors

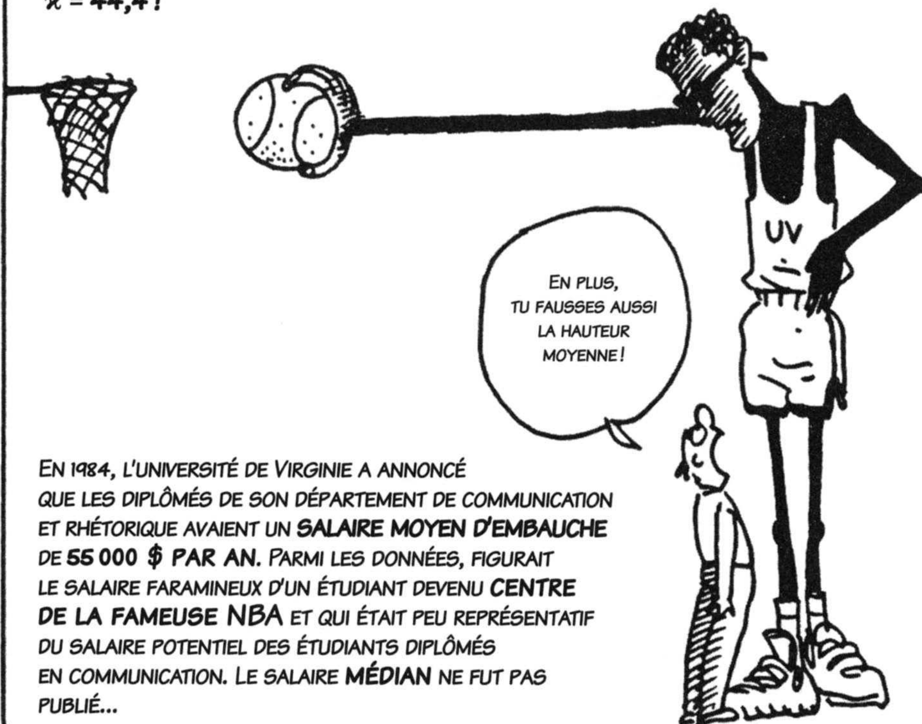
$$\frac{x_{46} + x_{47}}{2} = \frac{145 + 145}{2}$$

$$= \mathbf{145} \textbf{ livres} \text{ (environ 65,77 kg)}$$

Il y a exactement 46 données inférieures et 46 données supérieures à 145 livres.

9 : 5
10 : 288
11 : 002556688
12 : 00012355555
13 : 0000013555688
14 : 0000255**55**8
15 : 0000000000355555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5

Pourquoi utilisons-nous deux mesures différentes de tendance centrale? En fait, elles n'ont pas les mêmes propriétés. Par exemple, la **médiane** n'est pas sensible aux données extrêmes qui sont atypiques par rapport aux autres observations. Supposons que dans notre petit échantillon de téléspectateurs, une personne regarde la télévision 200 heures par semaine. Si nos données sont $\{3, 5, 5, 7, 200\}$, la médiane est encore 7 alors que la moyenne est maintenant $\bar{x} = \mathbf{44{,}4}$!

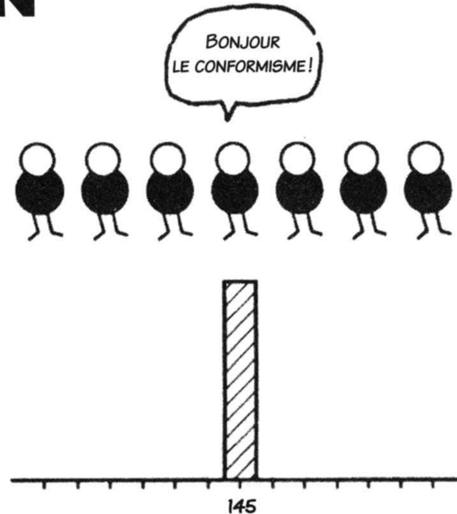


En 1984, l'Université de Virginie a annoncé que les diplômés de son département de communication et rhétorique avaient un **salaire moyen d'embauche de 55 000 $ par an**. Parmi les données, figurait le salaire faramineux d'un étudiant devenu **centre de la fameuse NBA** et qui était peu représentatif du salaire potentiel des étudiants diplômés en communication. Le salaire **médian** ne fut pas publié...

Mesures de DISPERSION

(ou de variabilité)

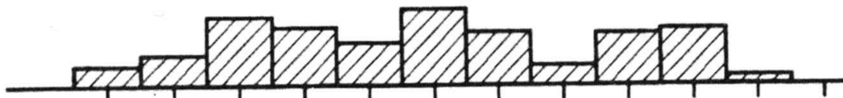
Outre le calcul de tendance centrale d'un ensemble de données, nous souhaiterions aussi connaître la **dispersion** des données (de combien les données s'éloignent du centre). Par exemple, si les étudiants pesaient tous **exactement** 145 livres, il n'y aurait aucune variabilité. Numériquement, la variance serait égale à **zéro** et l'histogramme serait très « léger ».



Mais si beaucoup d'étudiants étaient très maigres et/ou d'autres très lourds, nous aurions évidemment une grande variabilité – par exemple, si l'équipe de football américain faisait partie de l'échantillon...



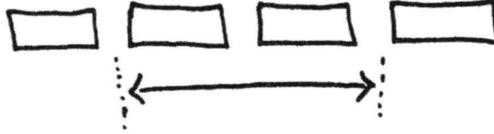
L'histogramme serait plus étendu, et il ressemblerait à cela :



À NOUVEAU, IL Y A PLUSIEURS FAÇONS DE DÉFINIR LA DISPERSION. L'UNE D'ELLES EST

L'ÉTENDUE INTERQUARTILE.

L'IDÉE EST DE DIVISER LES DONNÉES EN QUATRE GROUPES ET DE MESURER LA DISTANCE ENTRE LES DEUX GROUPES EXTRÊMES.



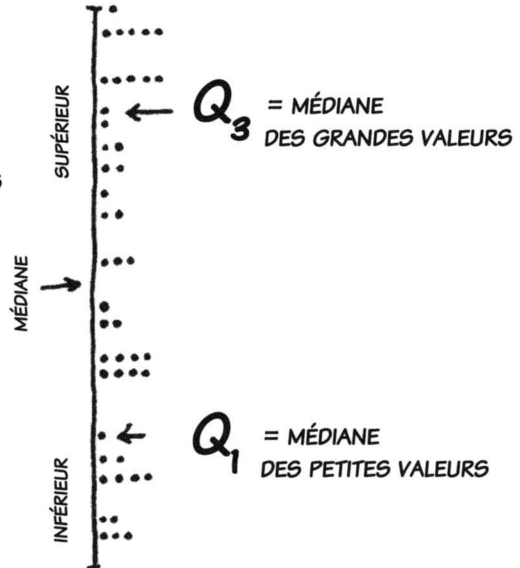
VOICI LA RECETTE :

1) ORDONNER LES DONNÉES PAR ORDRE CROISSANT.

2) DIVISER LES DONNÉES EN UN GROUPE EN DESSOUS ET UN GROUPE AU-DESSUS DE LA MÉDIANE (SI LA MÉDIANE CORRESPOND À UNE OBSERVATION, LA METTRE DANS CHAQUE GROUPE).

3) CHERCHER LA MÉDIANE DU GROUPE INFÉRIEUR. IL S'AGIT DU PREMIER QUARTILE NOTÉ Q_1.

4) CHERCHER LA MÉDIANE DU GROUPE SUPÉRIEUR. IL S'AGIT DU TROISIÈME QUARTILE NOTÉ Q_3.



MAINTENANT, L'ÉTENDUE INTERQUARTILE EIQ EST LA DISTANCE, OU LA DIFFÉRENCE, ENTRE CES DEUX QUARTILES :

$$EIQ = Q_3 - Q_1$$

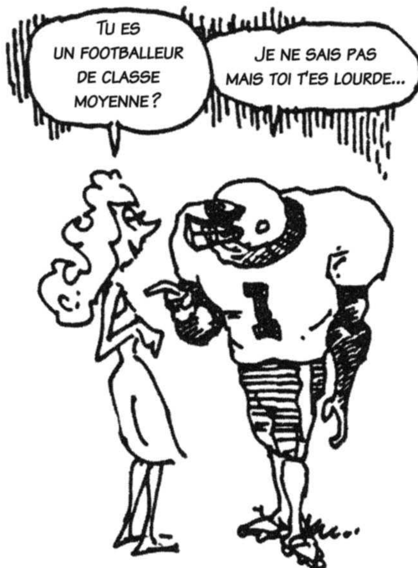
VOICI LES DONNÉES DE POIDS AVEC LES MÉDIANES DES POIDS INFÉRIEURS ET SUPÉRIEURS MIS EN ÉVIDENCE :

9 : 5
10 : 288
11 : 002556688 ← Q_1
12 : 00012355555
13 : 0000013555688 ← MÉDIANE
14 : 000025555 58
15 : 000000000035555555557 ← Q_3
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5

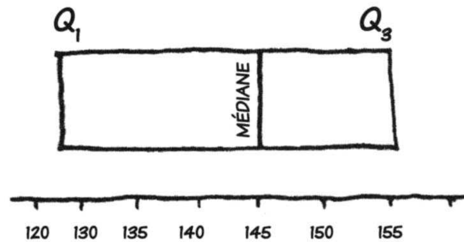
ET ON VOIT QUE

EIQ = 156 − 125
= 31 LIVRES

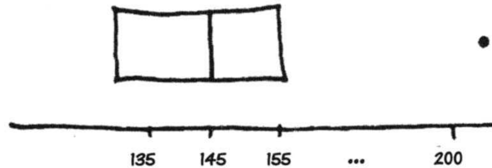
AINSI, IL S'AGIT DE LA DIFFÉRENCE ENTRE LA MÉDIANE DES POIDS ÉLEVÉS ET CELLE DES POIDS LÉGERS.



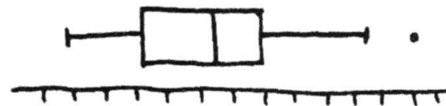
JOHN TUKEY A INVENTÉ UNE AUTRE FAÇON DE REPRÉSENTER L'EIQ, QUE L'ON APPELLE UNE **BOÎTE À PATTES.** LES LIMITES DE LA BOÎTE SONT LES QUARTILES Q_1 ET Q_3. ON TRACE LA MÉDIANE À L'INTÉRIEUR DE LA BOÎTE.



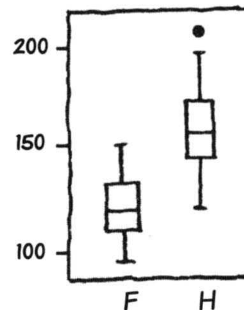
SI UN POINT EST À PLUS DE 1,5 EIQ DES LIMITES DE LA BOÎTE, ALORS C'EST UNE **VALEUR EXTRÊME.** ON LES REPRÉSENTE INDIVIDUELLEMENT PAR DES POINTS.



ENFIN, ON AJOUTE LES « PATTES » JUSQU'AUX DERNIÈRES OBSERVATIONS QUI NE SONT PAS DES VALEURS EXTRÊMES (C'EST-À-DIRE À MOINS DE 1,5 EIQ).



LES BOÎTES À PATTES SONT EXTRÊMEMENT UTILES POUR FAIRE RESSORTIR LES DIFFÉRENCES ENTRE DES GROUPES. VOICI LES DEUX REPRÉSENTATIONS POUR LES ÉTUDIANTS HOMMES ET FEMMES.



La mesure standard de variabilité est

L'ÉCART-TYPE.

Contrairement à l'EIQ, qui est basée sur les quartiles, l'écart-type mesure les écarts à la **moyenne.** On peut en gros retenir qu'il s'agit de l'écart moyen des données par rapport à la moyenne $\bar{x}$.

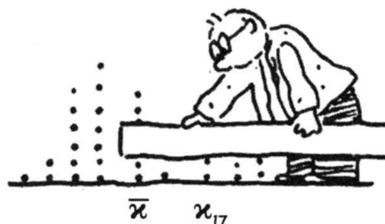


Sauf qu'en réalité, on utilise les **carrés** de ces écarts. Ainsi, le carré de l'écart entre x_i et $\bar{x}$ étant $(x_i - \bar{x})^2$, on obtient :

$$\text{moyenne des carrés des écarts} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Pour des raisons techniques, on utilise $n - 1$ au dénominateur plutôt que n, et on définit alors la **variance d'échantillon** s^2 comme :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$



Pour l'échantillon $\{3, 5, 7, 7, 38\}$, on avait $\bar{x} = 12$ et $n = 5$, on calcule alors la variance :

$$s^2 = \frac{(3-12)^2 + (5-12)^2 + (7-12)^2 + (7-12)^2 + (38-12)^2}{(5-1)}$$

$$= \frac{81 + 49 + 25 + 25 + 676}{4}$$

$$= 214$$



UNE MESURE DE DISPERSION DOIT ÊTRE DE MÊME UNITÉ QUE LES DONNÉES ORIGINALES. MALHEUREUSEMENT, DANS L'EXEMPLE DES POIDS, LA VARIANCE s^2 EST EN LIVRES AU **CARRÉ...** OUPS !





IL EST ÉVIDENT QU'IL SUFFIT DE PRENDRE LA RACINE CARRÉE, CE QUE NOUS FAISONS POUR DÉFINIR :

L'ÉCART-TYPE $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

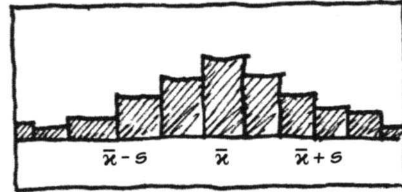
CE QUI DONNE POUR NOTRE ÉCHANTILLON DE DONNÉES :

$s = \sqrt{214} = 14{,}63$



MÊME POUR UN FAIBLE NOMBRE DE DONNÉES, LES CALCULS PEUVENT ÊTRE FASTIDIEUX ! HEUREUSEMENT AUJOURD'HUI, IL SUFFIT D'APPUYER SUR LE BOUTON D'UNE CALCULATRICE OU DE CONSULTER LE RAPPORT DE DONNÉES GÉNÉRÉ PAR UN LOGICIEL INCLUANT UN PACK STATISTIQUE.

Propriétés de $\overline{x}$ et s



LA MOYENNE ET L'ÉCART-TYPE RÉSUMENT TRÈS BIEN LES PROPRIÉTÉS DES **HISTOGRAMMES SYMÉTRIQUES SANS VALEURS EXTRÊMES,** C'EST-À-DIRE LES HISTOGRAMMES EN FORME DE MONTICULE.

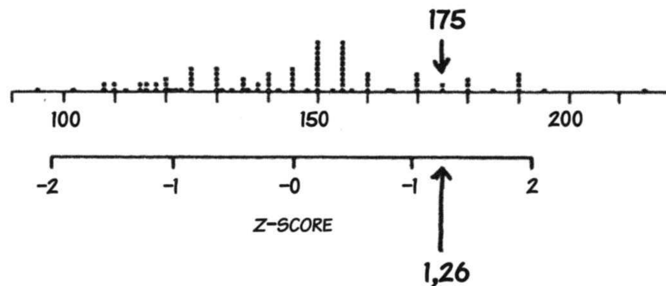


IL EST SOUVENT UTILE DE SAVOIR À **COMBIEN D'ÉCARTS-TYPES** SE TROUVE UNE DONNÉE PAR RAPPORT À LA MOYENNE. ON DÉFINIT ALORS LE **Z-SCORE,** OU VARIABLE CENTRÉE RÉDUITE, COMME LA DISTANCE D'UNE OBSERVATION À LA MOYENNE PAR **ÉCART-TYPE.**

$$z_i = \frac{x_i - \overline{x}}{s}$$ POUR CHAQUE i.



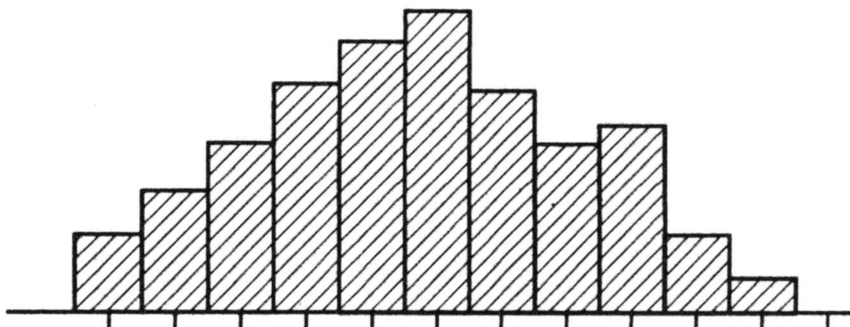
UN Z-SCORE DE + 2 SIGNIFIE QUE L'OBSERVATION EN QUESTION EST À **DEUX ÉCARTS-TYPES AU-DESSUS DE LA MOYENNE.** POUR L'ÉCHANTILLON DE POIDS ($\overline{x}$ = 145,2 ET s = 23,7), ON PEUT REPRÉSENTER LES DONNÉES INITIALES SUR L'AXE DES POIDS EN LIVRES ET LEUR Z-SCORE.



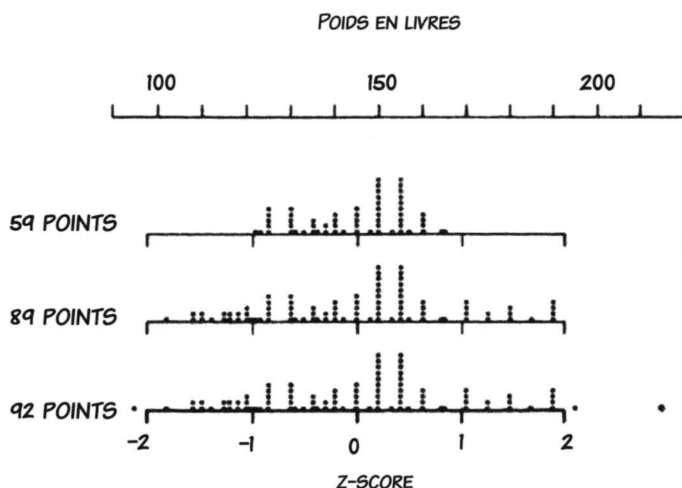
UN ÉTUDIANT QUI PÈSE 175 LIVRES A UN Z-SCORE DE $\frac{175 - 145,2}{23,7}$ = **1,26**

Une RÈGLE EMPIRIQUE :

POUR LA PLUPART DES ENSEMBLES DE DONNÉES DISTRIBUÉES SYMÉTRIQUEMENT EN FORME DE MONTICULE : APPROXIMATIVEMENT **68 %** DES DONNÉES SONT À UN ÉCART-TYPE DE LA MOYENNE ET **95 %** DES DONNÉES SONT À **DEUX** ÉCARTS-TYPES DE LA MOYENNE.



POUR LES POIDS DES ÉTUDIANTS, NOS DONNÉES VÉRIFIENT RELATIVEMENT LA RÈGLE : **64 %** (= 59/92) DES POIDS SONT À UN ÉCART-TYPE DE LA MOYENNE ET **97 %** DES POIDS SONT À DEUX ÉCARTS-TYPES DE LA MOYENNE.





ET MAINTENANT, APRÈS CES CALCULS NUMÉRIQUES, UN PEU DE REPOS !

ON A BEAUCOUP AVANCÉ DANS CE CHAPITRE ! À PARTIR D'UN TAS DÉSORGANISÉ DE NOMBRES, NOUS AVONS :

1) TROUVÉ DIFFÉRENTES FAÇONS DE LES REPRÉSENTER VISUELLEMENT ;

2) EXAMINÉ DEUX CONCEPTS DIFFÉRENTS DE TENDANCE CENTRALE : LA MÉDIANE ET LA MOYENNE ;

3) MESURÉ LA DISPERSION DES DONNÉES AUTOUR DU CENTRE DE DEUX FAÇONS DIFFÉRENTES ;

4) RENCONTRÉ DES HISTOGRAMMES EN FORME DE MONTICULE, ET DÉFINI Z, UNE VARIABLE QUI INDIQUE À COMBIEN D'ÉCARTS-TYPES ON SE TROUVE PAR RAPPORT À LA MOYENNE.

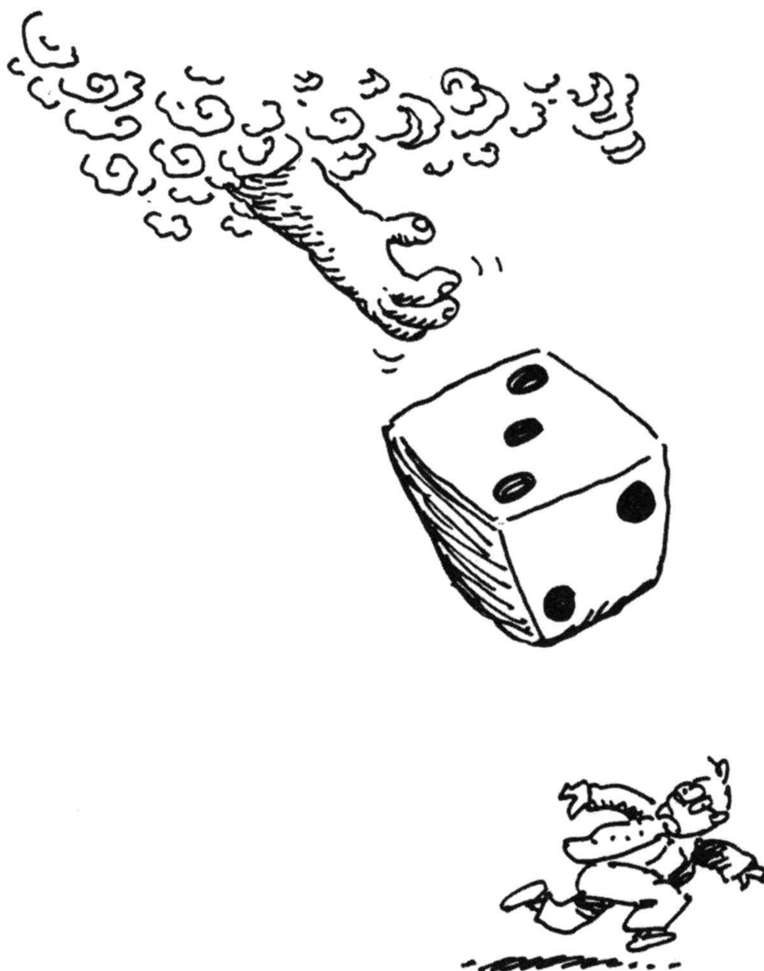


MAINTENANT, AFIN D'EXPLORER LE COMPORTEMENT DES DONNÉES DE FAÇON PLUS PRÉCISE, NOUS ALLONS FAIRE UN PETIT DÉTOUR DANS LE DOMAINE DE L'**ALÉATOIRE...** UN MONDE OÙ TOUT FONCTIONNE **TOUJOURS** À MERVEILLE SUR LE LONG TERME, ET OÙ LA SEULE LOI QUI TIENNE EST LA **LOI DU CASINO.**



Chapitre 3
LES PROBABILITÉS

RIEN N'EST SÛR ET CERTAIN DANS LA VIE. DANS TOUT CE QUE NOUS FAISONS, NOUS ÉVALUONS LES CHANCES DE SUCCÈS OU D'ÉCHEC, DU BUSINESS À LA MÉDECINE EN PASSANT PAR LA MÉTÉOROLOGIE. MAIS DEPUIS LES DÉBUTS DE L'AVENTURE HUMAINE, LE CALCUL DES **PROBABILITÉS,** QUI CONSISTE À ÉTUDIER FORMELLEMENT LES LOIS DU HASARD, AVANT TOUT, A ÉTÉ UTILISÉ DANS LE DOMAINE DES **JEUX D'ARGENT**.



Personne ne sait quand les jeux d'argent sont apparus dans l'histoire. Cela remonte au moins au temps de l'Égypte ancienne où des hommes et des femmes se divertissaient avec des **astragales** (os à quatre faces prélevés sur les pieds des animaux).



L'empereur romain **Claude Ier** (10 av. J.-C. – 54 apr. J.-C.) a écrit le premier traité sur les jeux de hasard. Malheureusement son livre *Comment gagner aux dés* a été perdu.



Les dés modernes ont été populaires au Moyen Âge puis à la Renaissance. Au XVIIe siècle, le **chevalier de Méré (1607-1684)** formula une énigme mathématique :



Le chevalier raisonna sur le nombre moyen de succès et il en déduisit que, dans les deux cas, ce nombre était le même :

Probabilité d'un six = $\frac{1}{6}$

Nombre moyen en 4 lancers = $4(1/6) = \frac{2}{3}$

Probabilité d'un double six = $\frac{1}{36}$

Nombre moyen en 24 lancers = $24(1/36) = \frac{2}{3}$

Alors, pourquoi **perdait-il plus souvent** avec le second pari ?



Méré posa la question à son ami, le génial **Blaise Pascal** (1623-1662).



Bien que Pascal ait cessé l'étude des mathématiques, jugée comme un divertissement d'ordre sexuel(!), il accepta de s'attaquer au problème de Méré.

Pascal écrivit à son collègue de génie **Pierre de Fermat (1601-1665)**, et après quelques échanges épistolaires, ils posèrent les bases de la théorie des probabilités sous sa forme moderne – excepté bien sûr cette version en bande dessinée.



Définitions préliminaires

De même qu'un parieur joue, de même nous allons jouer au scientifique en observant les résultats :

Une **expérience aléatoire** est un procédé qui permet d'observer un résultat, ou un événement, déterminé par un aléa.

Les **événements élémentaires** correspondent à tous les résultats possibles de l'expérience aléatoire.

L'**espace échantillon** est l'ensemble ou la collection de tous les événements élémentaires.



Par exemple, si l'événement est un lancer de pièce, l'**expérience aléatoire** consiste à enregistrer le résultat.



Les **événements élémentaires** sont soit face, soit pile.

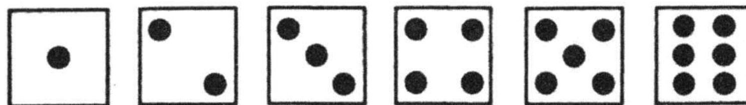


L'**espace échantillon** est l'ensemble suivant :

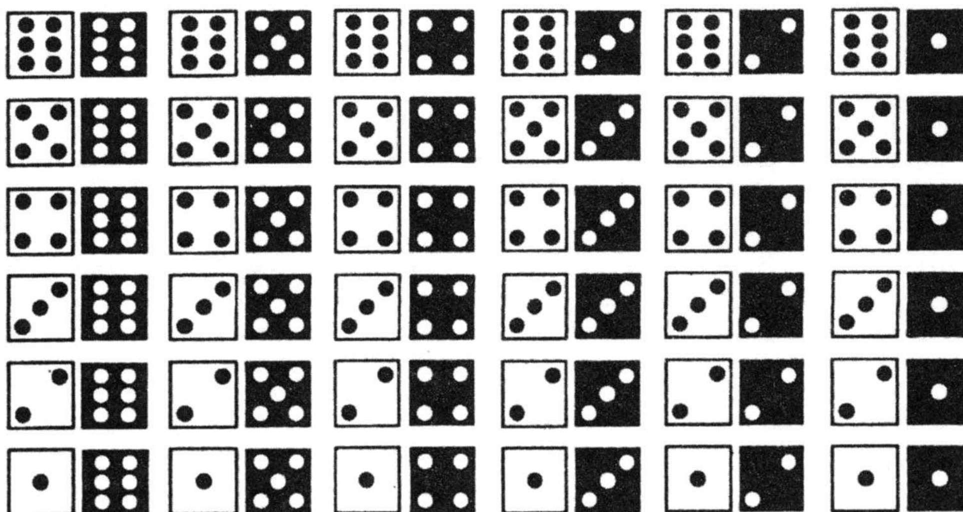
$$\{F, P\}$$



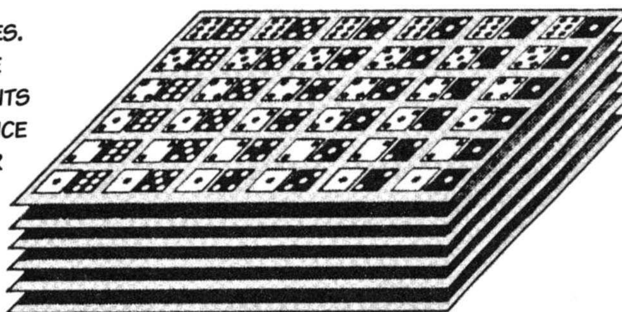
L'ESPACE ÉCHANTILLON D'UN LANCER D'**UN SEUL DÉ** EST UN PEU PLUS GRAND.



ET POUR LE LANCER D'UNE PAIRE DE DÉS, L'ESPACE ÉCHANTILLON RESSEMBLE À CECI (ON A PRIS UN DÉ BLANC ET UN DÉ NOIR POUR MIEUX DIFFÉRENCIER LES CAS) :



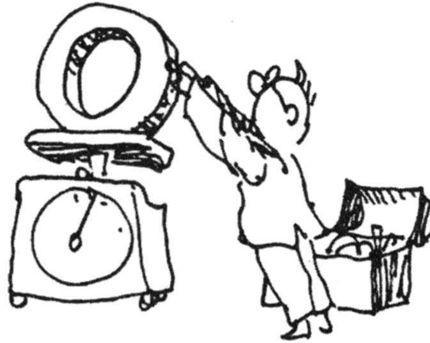
L'ESPACE ÉCHANTILLON COMPORTE **36** (6 × 6) ÉVÉNEMENTS ÉLÉMENTAIRES. AVEC TROIS DÉS, L'ESPACE COMPORTERAIT 216 ÉLÉMENTS COMME DANS CETTE MATRICE 6 × 6 × 6. ET ALORS, POUR **QUATRE** DÉS ?





AU BOUT D'UN MOMENT, IL FAUT ARRÊTER DE LISTER POUR COMMENCER À RÉFLÉCHIR...

IMAGINONS MAINTENANT UNE EXPÉRIENCE ALÉATOIRE AVEC n RÉSULTATS ÉLÉMENTAIRES NOTÉS O_1, O_2 ... O_n. ON VEUT ASSOCIER À CHAQUE RÉSULTAT UN **POIDS NUMÉRIQUE,** OU UNE **PROBABILITÉ** QUI MESURE LA VRAISEMBLANCE DE L'OCCURRENCE DE L'ÉVÉNEMENT. LA PROBABILITÉ DE O_i SE NOTE $P(O_i)$.



PAR EXEMPLE, DANS UN LANCER D'UNE PIÈCE (SANS TRICHER), LES CÔTÉS FACE ET PILE SONT ÉGALEMENT VRAISEMBLABLES. LA PROBABILITÉ DE CHACUN EST DE 0,5.

$$P(F) = P(P) = 0{,}5$$

CHAQUE ÉVÉNEMENT A LIEU UNE FOIS SUR DEUX. DEMANDEZ À UN **JOUEUR DE FOOTBALL!**

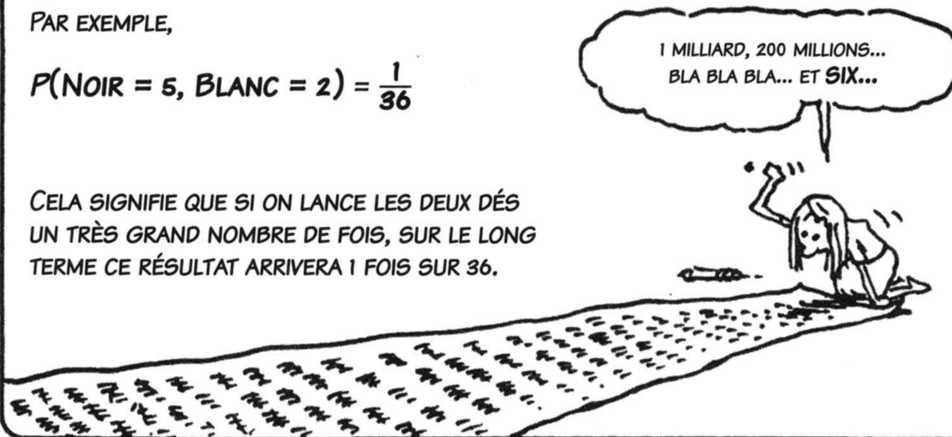


DANS UN LANCER DE **DEUX** DÉS, IL Y A **36** ÉVÉNEMENTS ÉLÉMENTAIRES. TOUS SONT AUSSI VRAISEMBLABLES, CHAQUE PROBABILITÉ EST DONC DE 1/36.

PAR EXEMPLE,

$$P(\text{NOIR} = 5, \text{BLANC} = 2) = \frac{1}{36}$$

CELA SIGNIFIE QUE SI ON LANCE LES DEUX DÉS UN TRÈS GRAND NOMBRE DE FOIS, SUR LE LONG TERME CE RÉSULTAT ARRIVERA 1 FOIS SUR 36.

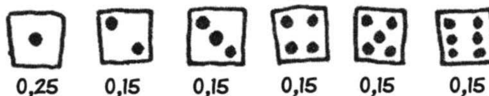


QUE SE PASSE-T-IL SI NOTRE JOUEUR **TRICHE** EN UTILISANT UN **DÉ PIPÉ ?** POUR SIMPLIFIER, SUPPOSONS QUE LA FACE 1 APPARAISSE **25 %** DU TEMPS (SUR LE LONG TERME).



L'**ESPACE ÉCHANTILLON** EST LE MÊME QUE CELUI D'UN DÉ NORMAL

$$\{1, 2, 3, 4, 5, 6\}$$



PAR CONTRE, LES PROBABILITÉS DIFFÈRENT. MAINTENANT $P(1) = 0{,}25$ ET LES PROBABILITÉS RESTANTES DOIVENT SE SOMMER À 0,75. SI 2, 3, 4, 5, ET 6 RESTENT ÉQUIPROBABLES, ALORS CHAQUE FACE AURA UNE PROBABILITÉ DE $0{,}15 = \frac{1}{5}(0{,}75)$.



EN GÉNÉRAL, LES ÉVÉNEMENTS ÉLÉMENTAIRES N'ONT PAS LA MÊME PROBABILITÉ.





MAINTENANT, QUE POUVONS-NOUS DIRE DES PROBABILITÉS $P(O_i)$ DANS UNE EXPÉRIENCE ALÉATOIRE QUELCONQUE. PREMIÈREMENT,

$$P(O_i) \geq 0$$

LES PROBABILITÉS NE SONT **JAMAIS NÉGATIVES.** UNE PROBABILITÉ DE ZÉRO VEUT DIRE QUE L'ÉVÉNEMENT N'AURA JAMAIS LIEU. UNE VALEUR STRICTEMENT INFÉRIEURE À ZÉRO N'AURAIT AUCUN SENS.



DEUXIÈMEMENT, S'IL EST **CERTAIN** QU'UN ÉVÉNEMENT AURA LIEU, NOUS LUI ASSIGNONS UNE PROBABILITÉ DE 1 (SUR LE LONG TERME, C'EST LA PROPORTION DE FOIS QU'IL SE PRODUIRA).



EN PARTICULIER, **LA PROBABILITÉ TOTALE DE L'ESPACE ÉCHANTILLON** DOIT ÊTRE ÉGALE À 1. SI L'ON CONDUIT L'EXPÉRIENCE, **QUELQUE CHOSE** ARRIVERA, FORCÉMENT !



EN COMBINANT LES DEUX, ON OBTIENT LES **PROPRIÉTÉS CARACTÉRISTIQUES DES PROBABILITÉS :**

$$P(O_i) \geq 0$$

UNE PROBABILITÉ EST **POSITIVE.**

$$P(O_1) + P(O_2) + \dots + P(O_n) = 1$$

LA PROBABILITÉ TOTALE DE TOUS LES ÉVÉNEMENTS ÉLÉMENTAIRES EST **UN.**





À LA FAÇON D'UN HABILE POLITICIEN, NOUS AVONS ÉVITÉ CERTAINES **QUESTIONS DÉPLAISANTES,** COMME :
A) QUE **SIGNIFIE** EXACTEMENT UNE PROBABILITÉ ?
ET B) **COMMENT** DÉFINIR LES PROBABILITÉS DES ÉVÉNEMENTS ÉLÉMENTAIRES ?

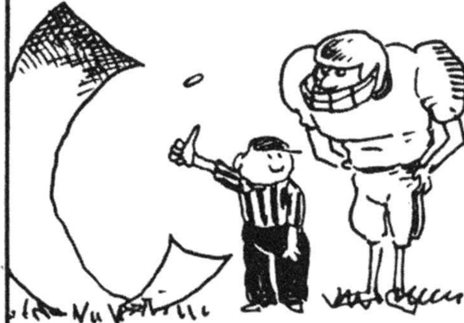


VOICI QUELQUES-UNES DES APPROCHES POSSIBLES :

LES PROBABILITÉS **classiques** : ELLES REPOSENT SUR DES IDÉES GÉNÉRALES SUR LES JEUX. L'HYPOTHÈSE FONDAMENTALE EST QUE LE JEU EST ÉQUITABLE ET QUE TOUS LES ÉVÉNEMENTS ÉLÉMENTAIRES ONT LA MÊME PROBABILITÉ.



LES **fréquences (relatives) :** QUAND UNE EXPÉRIENCE PEUT ÊTRE RÉPÉTÉE, ALORS LA PROBABILITÉ D'UN ÉVÉNEMENT EST LA PROPORTION D'OCCURRENCES DE L'ÉVÉNEMENT SUR LE LONG TERME.



LES PROBABILITÉS **personnelles :** LA PLUPART DES ÉVÉNEMENTS DE LA VIE NE SONT PAS **RÉPÉTABLES.** UNE PROBABILITÉ PERSONNELLE EST UNE ESTIMATION **INDIVIDUELLE ET PERSONNELLE** DE LA VRAISEMBLANCE D'UN ÉVÉNEMENT. SI UN PARIEUR ESTIME QU'UN CHEVAL A PLUS DE 50 % DE CHANCES DE GAGNER, IL MISERA SUR CE CHEVAL.



UN **OBJECTIVISTE** UTILISERA SOIT LA DÉFINITION CLASSIQUE, SOIT LES FRÉQUENCES COMME PROBABILITÉS. UN **SUBJECTIVISTE** OU UN **BAYÉSIEN*** APPLIQUERA LES LOIS FORMELLES DU HASARD POUR SES PROBABILITÉS PERSONNELLES OU LES VÔTRES.



*DE THOMAS BAYES (VOIR PAGE 47).

Les opérations de base

JUSQU'À PRÉSENT, NOUS AVONS DISCUTÉ UNIQUEMENT DES PROBABILITÉS DES ÉVÉNEMENTS ÉLÉMENTAIRES. EN THÉORIE, CELA SUFFIT POUR DÉCRIRE N'IMPORTE QUELLE EXPÉRIENCE ALÉATOIRE, MAIS DANS LA PRATIQUE, C'EST PEU MANIABLE. PAR EXEMPLE, L'OCCURRENCE D'UN JET DE SOMME 7 AVEC DEUX DÉS N'EST PAS UN ÉVÉNEMENT ÉLÉMENTAIRE. NOUS INTRODUISONS DONC UNE **NOUVELLE IDÉE.**



UN **ÉVÉNEMENT** EST UN ENSEMBLE D'ÉVÉNEMENTS ÉLÉMENTAIRES. LA PROBABILITÉ D'UN ÉVÉNEMENT EST LA SOMME DES PROBABILITÉS DES ÉVÉNEMENTS ÉLÉMENTAIRES DE L'ENSEMBLE. VOICI DES EXEMPLES D'ÉVÉNEMENTS LORSQU'ON LANCE DEUX DÉS :

DESCRIPTION DE L'ÉVÉNEMENT	ÉVÉNEMENTS ÉLÉMENTAIRES RÉALISANT L'ÉVÉNEMENT	PROBABILITÉS
A : SOMME DES DÉS = 3	{(1,2), (2,1)}	$P(A) = \frac{2}{36}$
B : SOMME DES DÉS = 6	{(1,5), (2,4), (3,3), (4,2), (5,1)}	$P(B) = \frac{5}{36}$
C : DÉ BLANC = 1	{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)}	$P(C) = \frac{6}{36}$
D : DÉ NOIR = 1	{(1,1), (2,1), (3,1), (4,1), (5,1), (6,1)}	$P(D) = \frac{6}{36}$



L'AVANTAGE D'UTILISER DES ÉVÉNEMENTS, ET PLUS SEULEMENT DES ÉVÉNEMENTS ÉLÉMENTAIRES, EST DE POUVOIR LES **COMBINER** POUR CRÉER DE NOUVEAUX ÉVÉNEMENTS EN UTILISANT DES OPÉRATEURS LOGIQUES. LES MOTS-CLÉS UTILISÉS SONT **ET**, **OU** ET **NON**.



AINSI, ÉTANT DONNÉ DEUX ÉVÉNEMENTS *E* ET *F*, ON PEUT CRÉER LES NOUVEAUX ÉVÉNEMENTS :

E **et** *F* : OÙ LES ÉVÉNEMENTS *E* ET *F* ONT LIEU EN MÊME TEMPS.

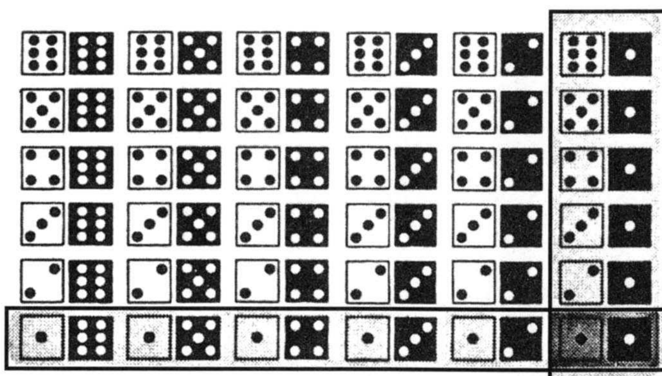
E **ou** *F* : OÙ SOIT *E* A LIEU, SOIT *F*, SOIT LES DEUX EN MÊME TEMPS.

non *E* : OÙ L'ÉVÉNEMENT *E* N'A PAS LIEU.

EN COMBINANT NOS DÉFINITIONS DE BASE SUR LES PROBABILITÉS AVEC CES OPÉRATEURS LOGIQUES, NOUS OBTENONS DE PUISSANTES FORMULES POUR MANIPULER LES PROBABILITÉS.



REVENONS À L'EXEMPLE DES LANCERS DE DÉS. SI *C* EST L'ÉVÉNEMENT [DÉ BLANC = 1] ET *D* EST L'ÉVÉNEMENT [DÉ NOIR = 1] ALORS



C* ET *D EST L'INTERSECTION (LE CHEVAUCHEMENT) DES PARTIES HACHURÉES OÙ LES DEUX DÉS VALENT 1.

C* OU *D EST L'UNION DES PARTIES HACHURÉES OÙ AU MOINS L'UN DES DÉS VAUT 1.

CELA ILLUSTRE LA RÈGLE D'ADDITION : POUR TOUT ÉVÉNEMENT *E* ET *F*,

$$P(E \text{ OU } F) = P(E) + P(F) - P(E \text{ ET } F)$$

EN EFFET, LA SOMME $P(E) + P(F)$ **COMPTE DEUX FOIS** LES ÉVÉNEMENTS ÉLÉMENTAIRES DE *E* ET *F*, NOUS DEVONS DONC RETIRER CETTE QUANTITÉ QUI CORRESPOND À $P(E \text{ ET } F)$.

DANS L'EXEMPLE CITÉ PLUS HAUT,

$$P(C \text{ OU } D) = \frac{11}{36}$$

COMME VOUS POUVEZ LE VOIR EN COMPTANT LES ÉVÉNEMENTS ÉLÉMENTAIRES. DE MÊME,

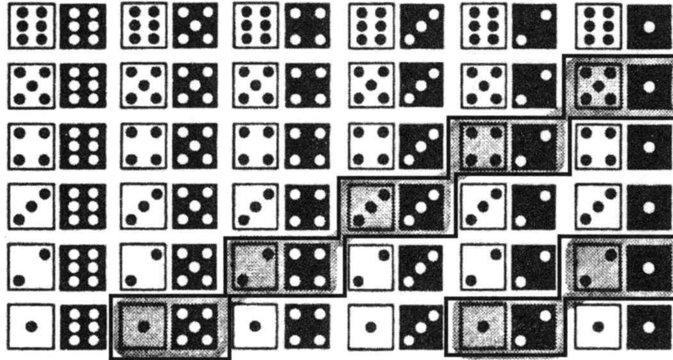
$$P(C \text{ ET } D) = \frac{1}{36}$$

ET ON CONFIRME ALORS LA FORMULE :

$$P(C) + P(D) - P(C \text{ ET } D)$$
$$= \frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36}$$
$$= P(C \text{ OU } D)$$



PARFOIS LE CHEVAUCHEMENT OU L'INTERSECTION **E ET F** EST VIDE, ET LES DEUX ÉVÉNEMENTS NE PARTAGENT AUCUN ÉVÉNEMENT ÉLÉMENTAIRE. DANS CE CAS, ON DIT QUE *E* ET *F* SONT **MUTUELLEMENT EXCLUSIFS** DE SORTE QUE $P(E \text{ ET } F) = 0$. DANS LE DESSIN SUIVANT, ON VOIT QUE LES ÉVÉNEMENTS *A* [SOMME DES DÉS = 3] ET *B* [SOMME DES DÉS = 6] SONT MUTUELLEMENT EXCLUSIFS.



POUR DES ÉVÉNEMENTS MUTUELLEMENT EXCLUSIFS, ON A UNE LOI SPÉCIALE D'ADDITION : SI *E* ET *F* SONT MUTUELLEMENT EXCLUSIFS ALORS

$$P(E \text{ OU } F) = P(E) + P(F)$$

ET ON PEUT VÉRIFIER $P(A \text{ OU } B) = \frac{7}{36} = \frac{2}{36} + \frac{5}{36} = P(A) + P(B)$

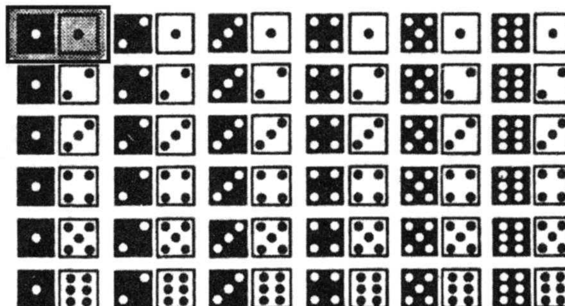
ET ENFIN UNE RÈGLE DE SOUSTRACTION : POUR TOUT ÉVÉNEMENT *E*,

$$P(E) = 1 - P(\text{NON } E)$$

CETTE RÈGLE EST TRÈS UTILE LORSQUE $P(\text{NON } E)$ EST PLUS FACILE À CALCULER QUE $P(E)$. PAR EXEMPLE, SOIT *E* L'ÉVÉNEMENT [PAS DE DOUBLE 1]. L'ÉVÉNEMENT *NON E*, [RÉUSSIR UN DOUBLE 1], A UNE PROBABILITÉ DE $P(\text{NON } E) = 1/36$.

AINSI

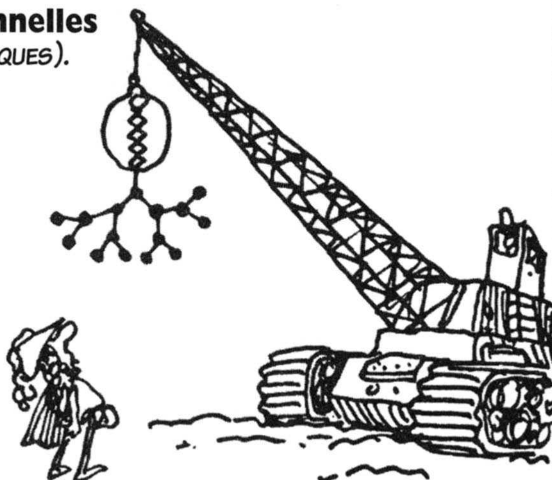
$$P(E) = 1 - P(\text{NON } E)$$
$$= 1 - \frac{1}{36}$$
$$= \frac{35}{36}$$





LES FORMULES QUE NOUS AVONS PRÉSENTÉES SONT, EN FAIT, ADAPTÉES POUR RÉSOUDRE LE PROBLÈME DE MÉRÉ. MAIS CE N'EST PAS SIMPLE ! (ON POURRAIT ESSAYER DE RÉPONDRE À UNE QUESTION PLUS FACILE : QUELLE EST LA PROBABILITÉ DE LANCER UN 6 SUR DEUX LANCERS DE DÉS ?) NOUS AVONS BESOIN DE **PLUS D'OUTILS** !

NOUS INTRODUISONS DONC LES **probabilités conditionnelles** (UN CONCEPT ESSENTIEL EN STATISTIQUES).



SUPPOSONS QUE NOUS MODIFIIONS NOTRE EXPÉRIENCE EN LANÇANT LE DÉ BLANC **AVANT** LE DÉ NOIR. QUELLE EST LA PROBABILITÉ DE L'ÉVÉNEMENT *A*, C'EST-À-DIRE QUE LA SOMME DES FACES SOIT 3 ?



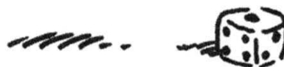
AVANT LE LANCEMENT DES DÉS, CETTE PROBABILITÉ EST

$$P(A) = \frac{2}{36}$$





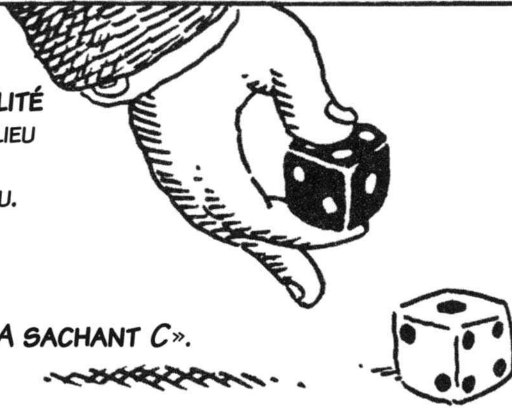
SUPPOSONS MAINTENANT QUE LE DÉ BLANC TOMBE SUR LE 1 (ÉVÉNEMENT *C*). QUELLE EST MAINTENANT LA PROBABILITÉ DE *A* ?



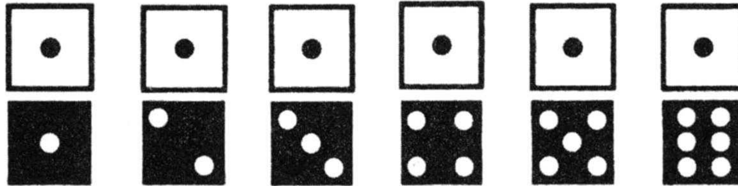
ON APPELLE CELA LA **PROBABILITÉ CONDITIONNELLE** QUE *A* AIT LIEU SOUS LA **CONDITION** QUE L'ÉVÉNEMENT *C* AIT DÉJÀ EU LIEU. ON L'ÉCRIT

$$P(A|C)$$

ET ON DIT « LA PROBABILITÉ DE *A* **SACHANT** *C* ».



AVANT DE JETER UN DÉ, L'ESPACE ÉCHANTILLON CONTIENT 36 ÉLÉMENTS. MAIS UNE FOIS QUE L'ÉVÉNEMENT *C* A EU LIEU, LE RÉSULTAT APPARTIENDRA À L'**ESPACE ÉCHANTILLON RÉDUIT VÉRIFIANT** *C*.



DANS L'ESPACE ÉCHANTILLON RÉDUIT DE SIX ÉVÉNEMENTS ÉLÉMENTAIRES, SEUL UN RÉSULTAT (1,2) DONNE UNE SOMME À 3. AINSI LA PROBABILITÉ CONDITIONNELLE EST DE 1/6.



EN GÉNÉRAL, POUR DÉTERMINER LA PROBABILITÉ CONDITIONNELLE $P(E|F)$, ON EXAMINE L'ÉVÉNEMENT ***E* ET *F*** EN TANT QUE PARTIE DE L'ESPACE ÉCHANTILLON RÉDUIT VÉRIFIANT *F*.



NOUS ALLONS TRANSCRIRE CELA EN DÉFINITION FORMELLE : LA **PROBABILITÉ CONDITIONNELLE DE E SACHANT F** EST :

$$P(E|F) = \frac{P(E \text{ ET } F)}{P(F)}$$



LA FORMULE VÉRIFIE DEUX RÉSULTATS INTUITIFS :

$P(E|E) = 1$ (SI E A EU LIEU, ALORS E EST CERTAIN).

LORSQUE E ET F SONT MUTUELLEMENT EXCLUSIFS

$P(E|F) = 0$ (SI F A EU LIEU, ALORS E EST IMPOSSIBLE).

EN RÉARRANGEANT LES TERMES, NOUS OBTENONS UNE **LOI MULTIPLICATIVE** :

$$P(E \text{ ET } F) = P(E|F)P(F)$$

CE QUE NOUS AIMERIONS SIMPLIFIER EN UNE LOI MULTIPLICATIVE « SPÉCIALE » SOUS LA CONDITION QUE $P(E|F) = P(E)$. CE SERAIT EXCELLENT !



L'INDÉPENDANCE et la loi multiplicative spéciale

DEUX ÉVÉNEMENTS ***E*** ET ***F*** SONT DITS **INDÉPENDANTS** L'UN DE L'AUTRE SI L'OCCURRENCE DE L'UN N'A **AUCUNE INFLUENCE** SUR L'OCCURRENCE DE L'AUTRE. PAR EXEMPLE, LE RÉSULTAT DU LANCER D'UN DÉ N'A AUCUNE INCIDENCE SUR LE RÉSULTAT DE L'AUTRE DÉ (À MOINS QU'ILS NE SOIENT COLLÉS ENTRE EUX, OU MAGNÉTIQUEMENT RELIÉS, ETC.!).



EN TERMES DE PROBABILITÉ CONDITIONNELLE, CELA IMPLIQUE QUE $P(E) = P(E|F)$ OU DE FAÇON ÉQUIVALENTE $P(F) = P(F|E)$. LORSQUE E ET F SONT INDÉPENDANTS, NOUS OBTENONS LA **LOI MULTIPLICATIVE SPÉCIALE :**

$$P(E \text{ ET } F) = P(E)P(F)$$

VÉRIFIONS MAINTENANT L'INDÉPENDANCE DES DÉS EN UTILISANT LES FORMULES. ***C*** EST L'ÉVÉNEMENT [DÉ BLANC = 1] ET ***D*** EST L'ÉVÉNEMENT [DÉ NOIR = 1]. NOUS AVONS ALORS

$$P(C|D) = \frac{P(C \text{ ET } D)}{P(D)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} = P(C)$$

PAR CONTRE, SI LE DÉ BLANC VAUT 1, CELA AFFECTE **ÉVIDEMMENT** LA PROBABILITÉ QUE LA SOMME DES DÉS DONNE 3 !

$$P(A|C) = \frac{P(A \text{ ET } C)}{P(C)} = \frac{P(1,2)}{P(C)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} \neq P(A) = \frac{1}{18}$$

CES DEUX ÉVÉNEMENTS A ET C **NE SONT** DONC **PAS INDÉPENDANTS.**

AVANT D'ALLER PLUS LOIN, VOICI LA LISTE DES LOIS QUE NOUS AVONS OBTENUES :

LOI ADDITIVE :

$$P(E \text{ OU } F) = P(E) + P(F) - P(E \text{ ET } F)$$

LOI SPÉCIALE D'ADDITION, SI **E** ET **F** SONT MUTUELLEMENT EXCLUSIFS :

$$P(E \text{ OU } F) = P(E) + P(F)$$

LOI DE SOUSTRACTION :

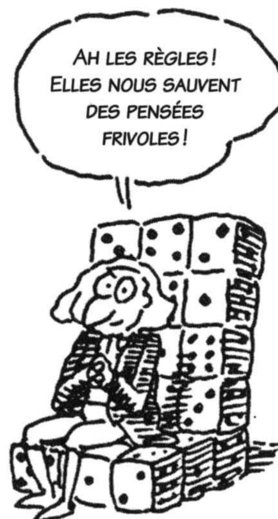
$$P(E) = 1 - P(\text{NON } E)$$

LOI MULTIPLICATIVE :

$$P(E \text{ ET } F) = P(E|F)P(F)$$

LOI SPÉCIALE DE MULTIPLICATION, SI **E** ET **F** SONT INDÉPENDANTS :

$$P(E \text{ ET } F) = P(E)P(F)$$



ET ENFIN LE PROBLÈME DE MÉRÉ... SOIT **E** L'ÉVÉNEMENT [OBTENIR AU MOINS UN SIX SUR QUATRE LANCERS DE DÉS], QUE VAUT **P(E) ?** IL S'AGIT D'UN ÉVÉNEMENT DONT L'ÉVÉNEMENT CONTRAIRE EST PLUS SIMPLE À CALCULER : **NON E** EST L'ÉVÉNEMENT [N'OBTENIR **AUCUN SIX SUR QUATRE LANCERS**].



SI A_i EST L'ÉVÉNEMENT [NE PAS OBTENIR DE SIX AU i^e LANCER], NOUS SAVONS QUE $P(A_i) = 5/6$. NOUS SAVONS AUSSI QUE LES LANCERS SONT INDÉPENDANTS. AINSI,

$$P(\text{NON } E) = P(A_1 \text{ ET } A_2 \text{ ET } A_3 \text{ ET } A_4)$$

LOI MULTIPLICATIVE →

$$= \left(\frac{5}{6}\right)^4 = \mathbf{0{,}482}$$

DONC

$$P(E) = 1 - P(\text{NON } E) = \mathbf{0{,}518}$$

Maintenant le second pari : soit **F** l'événement [obtenir un double 6 sur 24 lancers de deux dés]. À nouveau, **NON F** est plus simple à utiliser car c'est l'événement [n'obtenir **aucun double six**].



Si B_i est l'événement [pas de double six au i^e lancer] alors

$$\text{NON } F = B_1 \text{ et } B_2 \text{ et} \ldots B_{24}.$$

La probabilité de chaque B_i est :

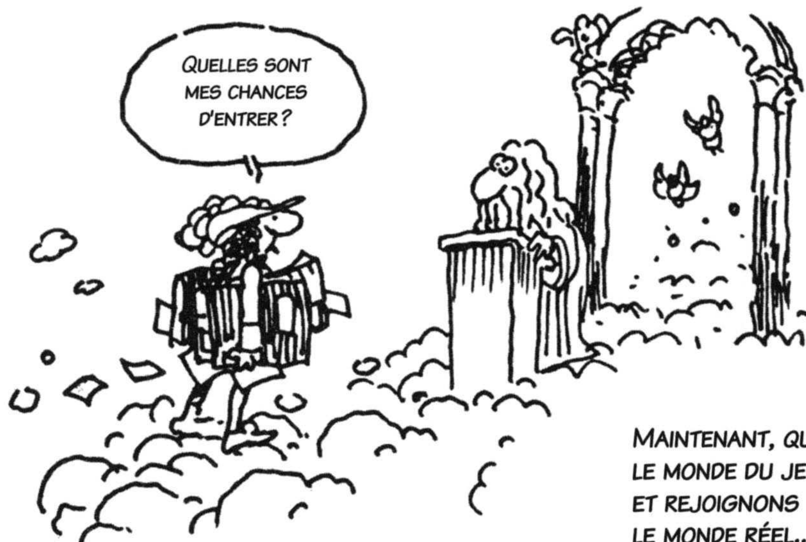
$P(B_i) = \left(\frac{35}{36}\right)$, donc

$$P(\text{NON } F) = \left(\frac{35}{36}\right)^{24} = 0{,}509$$

Par la règle de multiplication, on conclut que :

$$P(F) = 1 - P(\text{NON } F)$$
$$= 1 - 0{,}509 = \mathbf{0{,}491}$$

Méré avait dit à Pascal qu'il avait observé que l'événement **F** se produisait moins souvent que l'événement **E**, mais il restait perplexe sur l'explication à donner... Nous en concluons que Méré devait souvent faire le test et **en conserver soigneusement les résultats.**



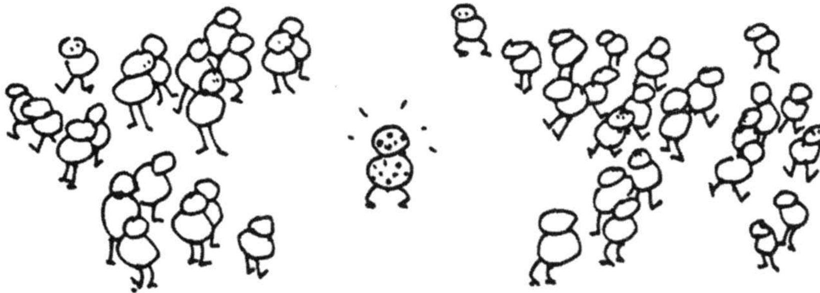
Maintenant, quittons le monde du jeu et rejoignons le monde réel...

LE THÉORÈME DE BAYES
et le cas des faux positifs

POUR UNE APPLICATION PLUS SÉRIEUSE DES PROBABILITÉS CONDITIONNELLES, EXAMINONS UN DOMAINE CONCERNANT UNE QUESTION DE VIE OU DE MORT.



SUPPOSONS QU'UNE MALADIE MORTELLE INFECTE UN INDIVIDU SUR 1 000 DANS UNE POPULATION DONNÉE...



ET SUPPOSONS QU'IL EXISTE UN BON TEST, MAIS IMPARFAIT, POUR DÉCELER LA MALADIE : SI UNE PERSONNE EST INFECTÉE, LE TEST EST POSITIF DANS **99 %** DES CAS. PAR AILLEURS, LE TEST PRODUIT AUSSI DES **FAUX POSITIFS** : POUR **2 %** DES PATIENTS SAINS LE TEST EST POSITIF. SI VOTRE TEST EST POSITIF, QUELLE PROBABILITÉ AVEZ-VOUS D'ÊTRE EFFECTIVEMENT INFECTÉ ?



NOUS AVONS DEUX ÉVÉNEMENTS À GÉRER :

A : LE PATIENT A LA MALADIE
B : LE PATIENT EST TESTÉ POSITIF

L'INFORMATION SUR L'EFFICACITÉ DU TEST PEUT S'ÉCRIRE :



$P(A) = 0{,}001$ UN PATIENT SUR 1 000 A LA MALADIE.

$P(B|A) = 0{,}99$ LA PROBABILITÉ D'UN TEST POSITIF, SACHANT QUE LE PATIENT EST MALADE, EST DE 0,99.

$P(B|\text{NON } A) = 0{,}02$ LA PROBABILITÉ D'UN FAUX POSITIF, SACHANT QUE LE PATIENT EST SAIN, EST DE 0,02.

ET LA QUESTION EST :

$P(A|B) = \text{QUOI ?}$ LA PROBABILITÉ D'AVOIR LA MALADIE, SACHANT QUE LE TEST EST POSITIF, EST INCONNUE.

COMME LE TRAITEMENT DE LA MALADIE A DE GRAVES EFFETS SECONDAIRES, LE MÉDECIN, SON AVOCAT ET L'AVOCAT DE SON AVOCAT FONT APPEL À JOE BAYES, CS (CONSULTANT STATISTICIEN) POUR OBTENIR UNE RÉPONSE. JOE UTILISE UN THÉORÈME PROUVÉ PAR SON ANCÊTRE, LE PASTEUR **THOMAS BAYES** (1701-1761).



JOE COMMENCE PAR FAIRE UNE MATRICE 2 × 2, QUI DIVISE L'ESPACE ÉCHANTILLON EN 4 CAS MUTUELLEMENT EXCLUSIFS. CHAQUE CAS EST UNE COMBINAISON ENTRE L'ÉTAT DU PATIENT ET LE RÉSULTAT DU TEST.

	A	NON A
B	A ET B	NON A ET B
NON B	A ET NON B	NON A ET NON B

INSCRIVONS LES PROBABILITÉS DE CHAQUE ÉVÉNEMENT DANS LA MATRICE :

	A	NON A	SOMME
B	P(A ET B)	P(NON A ET B)	P(B)
NON B	P(A ET NON B)	P(NON A ET NON B)	P(NON B)
SOMME	P(A)	P(NON A)	1

LES PROBABILITÉS EN FIN DE LIGNE ET DE COLONNE SONT OBTENUES EN SOMMANT LES LIGNES ET COLONNES.



MAINTENANT, CALCULONS :

$$P(A \text{ ET } B) = P(B|A)\,P(A) = (0{,}99)(0{,}001) = 0{,}00099$$

$$P(\text{NON } A \text{ ET } B) = P(B|\text{NON } A)\,P(\text{NON } A) = (0{,}02)(0{,}999) = 0{,}01998$$

CELA NOUS PERMET DE COMPLÉTER LA MATRICE :

	A	NON A	SOMME
B	0,00099	0,01998	0,02097
NON B	P(A ET NON B)	P(NON A ET NON B)	P(NON B)
SOMME	0,001	0,999	1

NOUS POUVONS TROUVER LES PROBABILITÉS MANQUANTES PAR SOUSTRACTION DANS LES COLONNES PUIS, EN ADDITIONNANT DANS LES LIGNES.

LA MATRICE FINALE EST :

	A	NON A	SOMME
B	0,00099	0,01998	0,02097
NON B	0,00001	0,97902	0,97903
SOMME	0,001	0,999	1
	P(A)	P(NON A)	

NOUS POUVONS ALORS EN DÉDUIRE :

$$P(A|B) = \frac{P(A \text{ ET } B)}{P(B)} = \frac{0,00099}{0,02097} = \mathbf{0,0472}$$

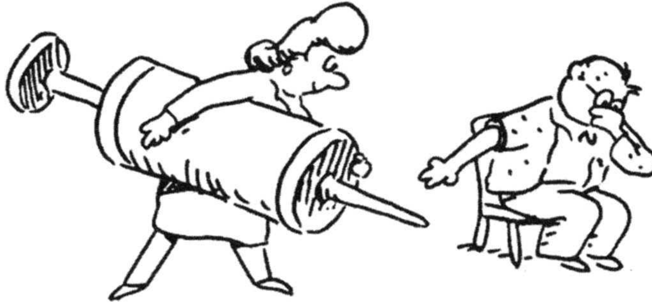
MALGRÉ L'APPARENTE FIABILITÉ DU TEST, **MOINS DE 5 %** DE CEUX QUI SONT TESTÉS POSITIFS SONT VÉRITABLEMENT MALADES ! ON APPELLE CELA LE **PARADOXE DES FAUX POSITIFS.**



CE TABLEAU EXPLIQUE LE PROBLÈME AVEC 1000 PATIENTS. EN MOYENNE, SEULEMENT 21 PERSONNES SERONT TESTÉES POSITIVES – ET SEULEMENT **UNE** PERSONNE SERA EFFECTIVEMENT MALADE ! IL Y AURA DONC 20 FAUX POSITIFS, CELA PROVIENT DU FAIT QUE **LE GROUPE SAIN EST DE LOIN LE PLUS IMPORTANT.**

	MALADES	BIEN PORTANTS	TOTAL
TESTS POSITIFS	1	20	21
TESTS NÉGATIFS	0	979	979
	1	999	1000

QUE DOIT FAIRE LE MÉDECIN ? JOE BAYES LUI CONSEILLE DE NE PAS COMMENCER LE TRAITEMENT SUR LA SEULE BASE DU TEST. LE TEST RESTE INFORMATIF : AVEC UN RÉSULTAT POSITIF, LA PROBABILITÉ QUE LE PATIENT SOIT MALADE PASSE DE 1 POUR 1000 À 1 POUR 21, MAIS DANS CE CAS LE MÉDECIN DEVRA FAIRE DE **NOUVEAUX TESTS.**



JOE BAYES TOUCHE SON CHÈQUE DE CONSULTANT AVANT D'ADMETTRE QUE TOUTE SON ÉTUDE POUVAIT SE RÉSUMER EN UNE SEULE FORMULE APPELÉE **THÉORÈME DE BAYES.**

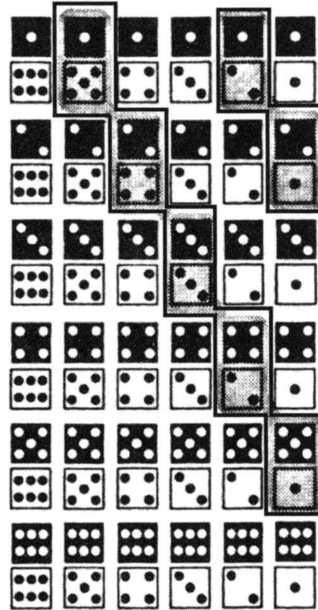
$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{NON } A)P(B|\text{NON } A)}$$



LA FORMULE EXPRIME $P(A|B)$ À PARTIR DE $P(A)$ ET DES DEUX PROBABILITÉS CONDITIONNELLES $P(B|A)$ ET $P(B|\text{NON } A)$. ON LA DÉMONTRE FACILEMENT EN NOTANT QUE LA FRACTION DE DROITE PEUT S'EXPRIMER COMME :

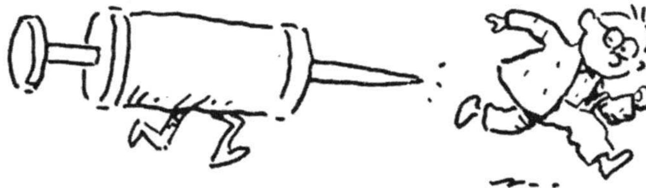
$$\frac{P(A \text{ ET } B)}{P(A \text{ ET } B) + P((\text{NON } A) \text{ ET } B)} = \frac{P(A \text{ ET } B)}{P(B)} = P(A|B)$$

DANS CE CHAPITRE, NOUS AVONS VU LES BASES DES PROBABILITÉS : LES DÉFINITIONS, L'ESPACE ÉCHANTILLON ET LES ÉVÉNEMENTS ÉLÉMENTAIRES, LES PROBABILITÉS CONDITIONNELLES ET CERTAINES FORMULES ESSENTIELLES POUR CALCULER LES PROBABILITÉS. NOUS AVONS ILLUSTRÉ CES IDÉES AVEC L'EXEMPLE D'UN LANCER DE DEUX DÉS. POUR UN JOUEUR MODERNE, LES PROBABILITÉS SONT UN OUTIL DE CHOIX.





ENFIN, AVEC L'EXEMPLE MÉDICAL, NOUS AVONS MONTRÉ COMMENT DES IDÉES ABSTRAITES POUVAIENT NOUS AIDER À PRENDRE DE BONNES DÉCISIONS DANS UN CONTEXTE D'**INFORMATION IMPARFAITE** ET DE **RISQUES RÉELS** – CE QUI EST LE BUT ULTIME DES STATISTIQUES.



MAIS CE N'EST QU'UN DÉBUT. POUR NOUS, LES PROBABILITÉS CONSTITUENT UN **OUTIL** (SANS CONTESTE ESSENTIEL) DANS L'ÉTUDE DES STATISTIQUES. DANS LES CHAPITRES QUI SUIVENT, NOUS ALLONS EXPLORER LA RELATION SUBTILE ENTRE LES PROBABILITÉS, LES VARIATIONS DANS LES DONNÉES STATISTIQUES, ET NOTRE CONFIANCE DANS L'INTERPRÉTATION DE NOS OBSERVATIONS.





Chapitre 4
LES VARIABLES ALÉATOIRES

DANS LE CHAPITRE 2, NOUS AVONS VU QUE L'OBSERVATION DE DONNÉES NUMÉRIQUES, COMME LE POIDS DES ÉTUDIANTS, PEUT ÊTRE REPRÉSENTÉE GRAPHIQUEMENT OU RÉSUMÉE EN TERMES DE TENDANCE CENTRALE, DE DISPERSION, DE VALEURS EXTRÊMES, ETC. DANS LE CHAPITRE 3, NOUS AVONS VU COMMENT DES PROBABILITÉS PEUVENT ÊTRE AFFECTÉES À DES RÉSULTATS D'UNE EXPÉRIENCE ALÉATOIRE.



SI NOUS IMAGINONS UNE EXPÉRIENCE ALÉATOIRE RÉPÉTÉE DE NOMBREUSES FOIS, NOUS NOUS ATTENDONS À CE QUE LES FRÉQUENCES DES RÉSULTATS OBSERVÉS TENDENT VERS LEURS PROBABILITÉS À TERME. LES PROBABILITÉS CONSTITUENT UN **MODÈLE** POUR LES EXPÉRIENCES DE LA VIE RÉELLE... ALORS, POURQUOI NE PAS FAIRE L'ÉTUDE D'UN MODÈLE À PARTIR DE L'ANALYSE DES DONNÉES DÉCRITES PAR CE MODÈLE ?

Le concept essentiel est celui de variable aléatoire, que nous écrivons avec un grand



X

Une variable aléatoire est définie comme le **résultat numérique** d'une **expérience aléatoire.**

Par exemple, imaginons qu'on sélectionne un étudiant au hasard de notre groupe d'étudiants. Il s'agit bien d'une expérience aléatoire. La **taille,** le **poids,** le **revenu familial,** le **résultat au bac** et la **moyenne générale** de l'étudiant sont des **variables numériques** décrivant les propriétés de l'étudiant tiré au hasard. Ce sont toutes des variables aléatoires.



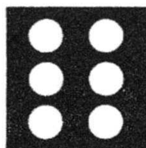
Un autre exemple : lancer deux pièces (l'expérience aléatoire) et noter le **nombre** de faces : 0, 1 ou 2.

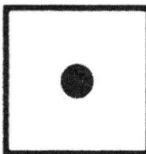
Résultat	PP	PF	FP	FF
x	0	1		2



Attention pour la notation, la variable s'écrit avec un **X** majuscule. Le x en minuscule représente une valeur particulière de X, ainsi $x = 2$ veut dire qu'il y a eu deux fois face.

Voici un autre exemple basé sur le lancer classique de deux dés. Soit **Y** la somme des points visibles sur les deux dés. Pour cette variable aléatoire, **Y** est un nombre compris entre 2 et 12.





Y = 7

Nous voulons maintenant examiner les **probabilités** des résultats. On écrit $P(X = x)$, ou simplement $p(x)$, la probabilité que la variable aléatoire **X** soit égale à x. Pour le cas de la variable aléatoire du lancer de deux pièces, nous avons le tableau suivant :

x	0	1	2
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

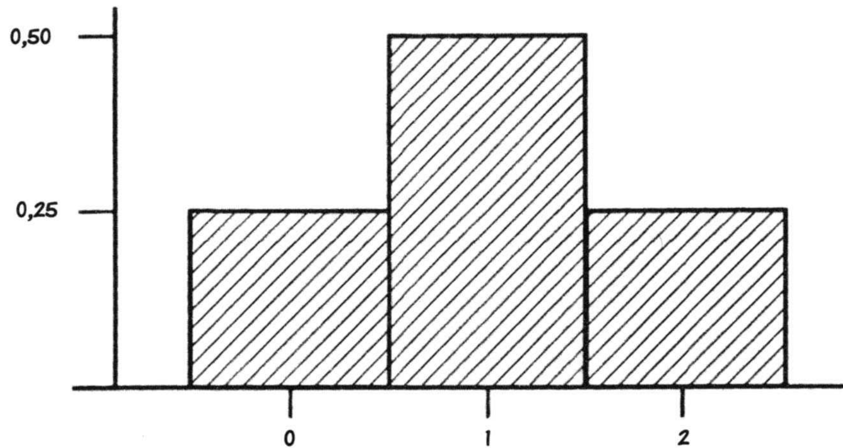
Ce tableau est appelé la **distribution de probabilités** de la variable aléatoire **X**.

Pour la variable aléatoire **Y** (somme du lancer de deux dés), la distribution de probabilités ressemble à ceci :

y	2	3	4	5	6	7	8	9	10	11	12
$P(Y = y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

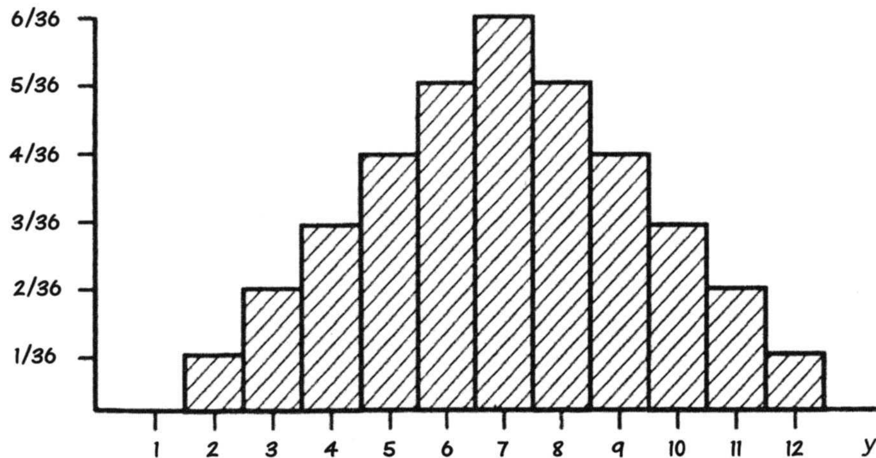


MAINTENANT DESSINONS DES GRAPHIQUES, OU DES **HISTOGRAMMES** REPRÉSENTANT LES DISTRIBUTIONS DE PROBABILITÉS. POUR CHAQUE VALEUR DE X, ON DESSINE UNE BARRE DE HAUTEUR $p(x)$.

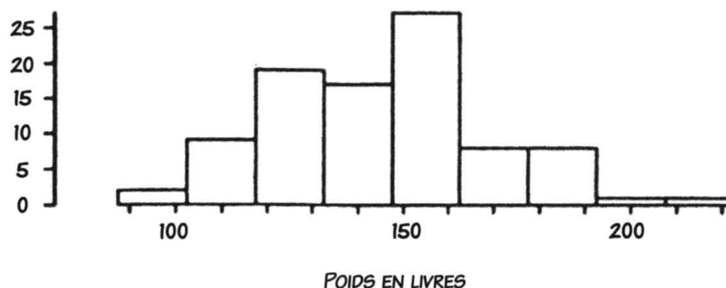


IL EST FACILE DE VOIR QUE L'AIRE TOTALE DE CES RECTANGLES EST 1. CHAQUE RECTANGLE A POUR BASE 1 ET UNE HAUTEUR DE $p(x)$. AINSI L'AIRE TOTALE EST ÉGALE À LA SOMME DES PROBABILITÉS DE TOUS LES RÉSULTATS POSSIBLES, C'EST-À-DIRE 1.

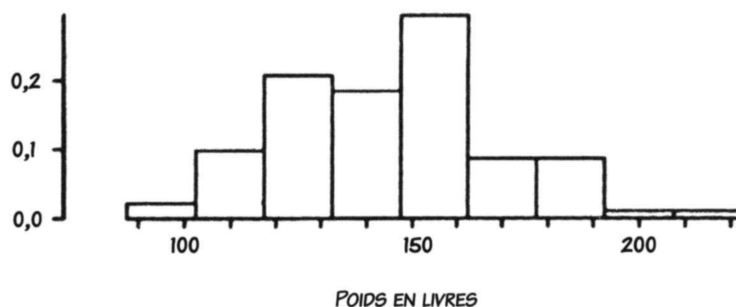
VOICI L'HISTOGRAMME DES PROBABILITÉS DE LA VARIABLE ALÉATOIRE Y : IL DÉCRIT LA DISTRIBUTION DE PROBABILITÉS DE LA SOMME DE DEUX DÉS.



POURQUOI APPELLE-T-ON CES GRAPHIQUES DES HISTOGRAMMES ? VOUS VOUS RAPPELEZ QUE DANS LE CHAPITRE 2, UN HISTOGRAMME ÉTAIT UN GRAPHIQUE QUI REPRÉSENTAIT LE NOMBRE DE DONNÉES OBSERVÉES DANS CHACUNE DES CLASSES.



À PARTIR DE CET HISTOGRAMME DES **EFFECTIFS**, ON A DÉFINI UN HISTOGRAMME DES **FRÉQUENCES** QUI REPRÉSENTE LA **PROPORTION** DE DONNÉES OBSERVÉES DANS CHAQUE CLASSE.



MAIS, VOUS VOUS SOUVENEZ QUE PAR DÉFINITION UNE PROBABILITÉ REPRÉSENTE LA FRÉQUENCE D'UN ÉVÉNEMENT SUR LE « LONG TERME ». SI L'ON RÉPÈTE L'EXPÉRIENCE ALÉATOIRE DE NOMBREUSES FOIS, L'HISTOGRAMME EN **FRÉQUENCE** DES RÉSULTATS OBSERVÉS DOIT ÊTRE SEMBLABLE À L'HISTOGRAMME DES **PROBABILITÉS** DE LA VARIABLE ALÉATOIRE.







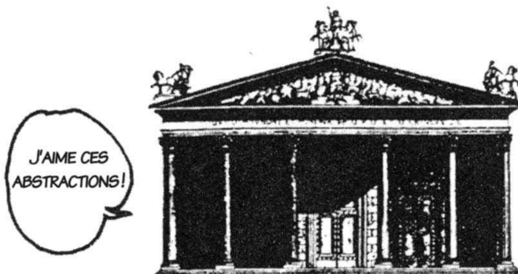
NOUS CONNAISSONS LA DISTRIBUTION DE PROBABILITÉS DE X ET NOUS SAVONS QUE LES RÉSULTATS DE L'EXPÉRIENCE RÉELLE CORRESPONDRONT APPROXIMATIVEMENT AUX PROBABILITÉS.
APRÈS 1000 LANCERS, LA LANCEUSE FAIT LE POINT SUR SES DONNÉES :

MODÈLE DE PROBABILITÉS		DONNÉES OBSERVÉES	
$p(x)$	x	n_x = NOMBRE D'OCCURRENCES	n_x/n = FRÉQUENCE
0,25	0	260	0,260
0,5	1	517	0,517
0,25	2	223	0,223

ET NOUS CONSTATONS QUE L'HISTOGRAMME DES PROBABILITÉS DE X CORRESPOND À UNE « FORME PURE » OU AU MODÈLE DE L'HISTOGRAMME EN FRÉQUENCE DES DONNÉES OBSERVÉES.



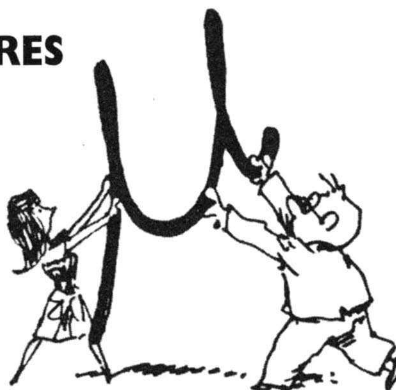
POUR POURSUIVRE L'ANALOGIE ENTRE FRÉQUENCE ET DONNÉES, NOUS ALLONS PARLER DE LA MOYENNE ET DE LA VARIANCE (OU DE L'ÉCART-TYPE) D'UNE DISTRIBUTION DE PROBABILITÉS...



ET AFIN DE NOUS RAPPELER QUE NOUS SOMMES DANS LE DOMAINE DE L'ABSTRACTION, NOUS MOBILISONS DES **LETTRES GRECQUES...**

MOYENNE et VARIANCE des VARIABLES ALÉATOIRES

ON UTILISE UNE TERMINOLOGIE ET DES SYMBOLES PARTICULIERS POUR DISTINGUER LES CARACTÉRISTIQUES DE DONNÉES OBSERVÉES DE CELLES DE PROBABILITÉS DE DISTRIBUTION.



LES CARACTÉRISTIQUES NUMÉRIQUES DE DONNÉES SONT APPELÉES CARACTÉRISTIQUES D'**ÉCHANTILLON,** ALORS QUE LES CARACTÉRISTIQUES D'UNE PROBABILITÉ DE DISTRIBUTION SONT APPELÉES CARACTÉRISTIQUES DU **MODÈLE** OU DE LA **POPULATION.** ON UTILISE DES LETTRES GRECQUES COMME μ (MU) POUR LA MOYENNE DE POPULATION ET σ (SIGMA MINUSCULE) POUR L'ÉCART-TYPE DE POPULATION (POUR UN ÉCHANTILLON DE DONNÉES, ON UTILISE LES LETTRES ROMAINES $\bar{x}$ ET s).

PARCE QUE LES ROMAINS ÉTAIENT LÉGERS POUR LA THÉORIE MAIS FORTS EN CONSTRUCTION... ET D'AUTRES CHOSES DE CE GENRE.

La moyenne échantillon a été définie par l'équation

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Certaines de ces données observées x_i peuvent avoir des valeurs égales. Dans l'exemple de la lanceuse de pièces, les seules valeurs possibles sont 0, 1 ou 2 et celle-ci a fait 1000 lancers. La valeur 0 est apparue 260 fois, la valeur 1 : 517 fois, et la valeur 2 : 223 fois.

La variable x va prendre toutes les valeurs possibles, notons n_x le nombre de données dont la valeur est x. On peut alors réécrire la formule :



$$\bar{x} = \frac{1}{n}\sum_{\text{tout } x}^{n} n_x x$$

ou

$$\bar{x} = \sum_{\text{tout } x} x \frac{n_x}{n}$$

Ah ! Mais n_x/n correspond à la **fréquence**... c'est-à-dire la « probabilité approximée... » soit le nombre qui tend vers $p(x)$... Donc par analogie nous obtenons la formule

$$\sum_{\text{tout } x} x p(x)$$



et définissons cela comme l'**espérance** (ou moyenne) de la **distribution de probabilité.**

Définition : LA **moyenne** D'UNE VARIABLE ALÉATOIRE X EST DÉFINIE PAR :

$$\mu = \sum_{\text{TOUT } x} x p(x)$$



ON L'APPELLE AUSSI ESPÉRANCE OU **VALEUR ESPÉRÉE** DE X, OU **E[X].** ON PEUT RETENIR QU'IL S'AGIT DE LA SOMME DE TOUTES LES VALEURS POSSIBLES PONDÉRÉE PAR LES PROBABILITÉS.

DANS L'EXPÉRIENCE DE LA LANCEUSE DE PIÈCES, ON PEUT COMPARER LA MOYENNE D'ÉCHANTILLON $\bar{x}$ AVEC LA MOYENNE DE POPULATION μ :

ÉCHANTILLON

x	$\frac{n_x}{n}$	$x\frac{n_x}{n}$
0	0,26	0,0
1	0,517	0,517
2	0,223	0,446
		0,963 = $\bar{x}$

MODÈLE

x	$p(x)$	$xp(x)$
0	0,25	0,0
1	0,5	0,5
2	0,25	0,5
		1 = μ

MAINTENANT, FAISONS LA MÊME CHOSE AVEC LA **VARIANCE.** PEUT-ÊTRE VOUS RAPPELEZ-VOUS LA FORMULE :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

CELA MESURE (PRESQUE) LA MOYENNE DES CARRÉS DES ÉCARTS À LA MOYENNE. COMME PRÉCÉDEMMENT, ON PEUT LA RÉÉCRIRE :

$$s^2 = \sum_{\text{TOUT } x}(x_i - \bar{x})^2 \frac{n_x}{n-1}$$



IL S'AGIT BIEN D'UNE SOMME PONDÉRÉE D'ÉCARTS AU CARRÉ, MIS À PART CE $n - 1$ AU DÉNOMINATEUR AU LIEU DE n... NOUS DÉFINISSONS DONC :

LA **variance** D'UNE VARIABLE ALÉATOIRE X COMME LA VALEUR ESPÉRÉE DES CARRÉS DES ÉCARTS À LA MOYENNE DE POPULATION :

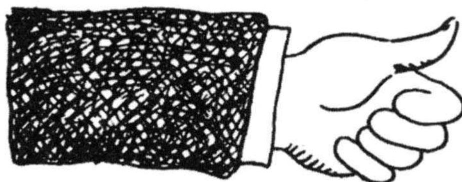
$$\sigma^2 = \sum_{\text{TOUT } x} (x - \mu)^2 p(x)$$

L'**écart-type** σ EST LA RACINE CARRÉE DE LA VARIANCE.



POUR TROUVER LA VARIANCE DU CAS DU LANCER DE 2 PIÈCES (POUR LEQUEL $\mu = 1$), ON UTILISE LE TABLEAU DE LA PAGE PRÉCÉDENTE.

x	$p(x)$	$(x - \mu)^2 p(x)$
0	0,25	$(0 - 1)^2(0,25) = 0,25$
1	0,5	$(1 - 1)^2(0,5) = 0$
2	0,25	$(2 - 1)^2(0,25) = 0,25$
		$0,50 = \sigma^2$



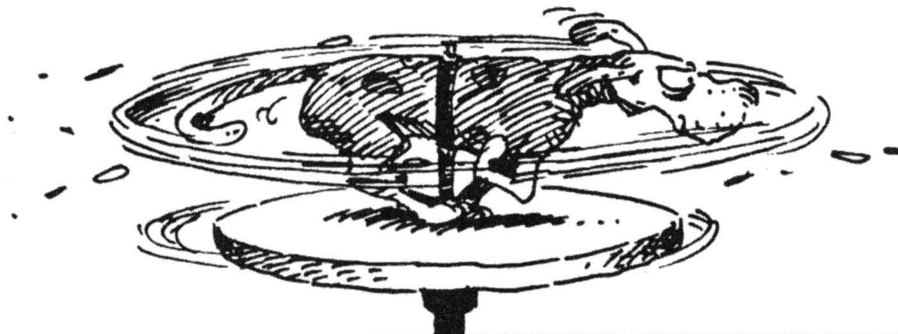
EN RÉSUMÉ : μ ET σ, LES MOYENNES ET ÉCARTS-TYPES DE POPULATION SONT DES CARACTÉRISTIQUES QUE L'ON CALCULE POUR DES **DISTRIBUTIONS** DE **PROBABILITÉS.** ELLES SONT ANALOGUES AUX MOYENNES $\bar{x}$ ET AUX ÉCARTS-TYPES s D'ÉCHANTILLON QUE L'ON CALCULE À PARTIR DE DONNÉES D'ÉCHANTILLON.

JUSQU'À PRÉSENT NOS EXEMPLES DE VARIABLES ALÉATOIRES ÉTAIENT DISCRETS. LES RÉSULTATS ÉTAIENT DES VALEURS ISOLÉES (OU « DISCRÈTES ») COMME DANS LE CHAPITRE 3, MAIS IL Y A AUSSI DES **variables aléatoires continues.**

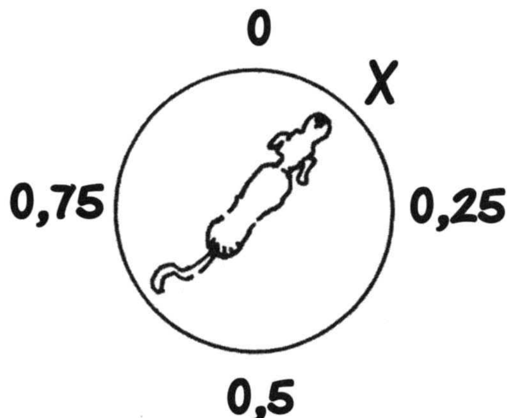


IMAGINONS UNE EXPÉRIENCE ALÉATOIRE OÙ CHAQUE RÉSULTAT AIT UNE PROBABILITÉ DE ZÉRO. AUTREMENT DIT, $p(x) = 0$ POUR TOUT x.

UN EXEMPLE SIMPLE EST CELUI D'UN POINTEUR SUR UNE ROUE (SI VOUS N'AVEZ PAS DE CHIEN DE CHASSE, UNE AIGUILLE FERA L'AFFAIRE). CE DERNIER PEUT S'ARRÊTER N'IMPORTE OÙ DANS LE CERCLE. SOIT X LA PROPORTION DE LA CIRCONFÉRENCE ATTEINTE PAR LE POINTEUR. LA VARIABLE ALÉATOIRE X PEUT PRENDRE N'IMPORTE QUELLE VALEUR ENTRE 0 ET 1, ET DONC UN ÉVENTAIL **INFINI** DE VALEURS.

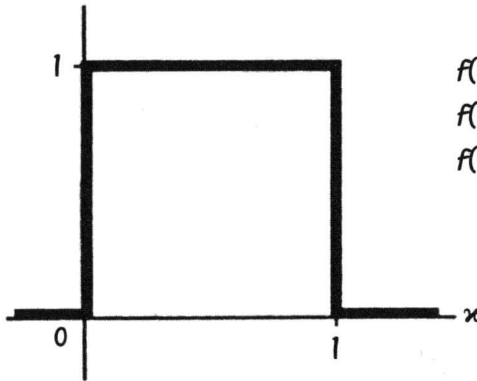


IL EST FACILE DE TROUVER LA PROBABILITÉ QUE X APPARTIENNE À UN **INTERVALLE** DONNÉ : PAR EXEMPLE, $P(0{,}25 \leq X \leq 0{,}75) = 0{,}5$ CAR IL S'AGIT DE LA MOITIÉ DU CERCLE. MAIS QU'EN EST-IL DE $P(X = 0{,}5)$? COMME X PEUT PRENDRE UN NOMBRE INFINI DE VALEURS ET QUE CHAQUE VALEUR EST ÉQUIPROBABLE, LA PROBABILITÉ QUE X SOIT **EXACTEMENT 0,5** (OU TOUTE AUTRE VALEUR) EST PRÉCISÉMENT NULLE.

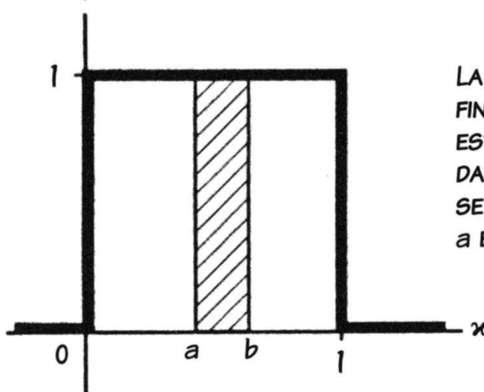


COMMENT REPRÉSENTER CELA ?
PAR ANALOGIE AVEC LE CAS DES PROBABILITÉS DISCRÈTES, IL FAUT VOIR LES PROBABILITÉS CONTINUES COMME L'**AIRE SITUÉE SOUS QUELQUE CHOSE.** DANS LE CAS DU POINTEUR DE LA ROUE, CE **QUELQUE CHOSE** RESSEMBLE À CECI :

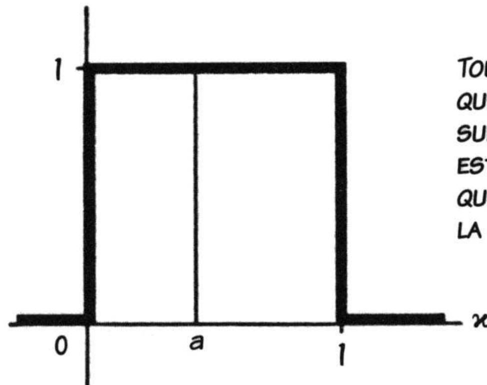




$f(x) = 0$ SI $x < 0$
$f(x) = 1$ SI $0 \leq x \leq 1$
$f(x) = 0$ SI $x > 1$



LA PROBABILITÉ QUE LE POINTEUR FINISSE SA COURSE ENTRE a ET b EST PRÉCISÉMENT L'AIRE SITUÉE DANS LA PARTIE HACHURÉE QUI SE TROUVE SOUS LA COURBE ENTRE a ET b. CETTE AIRE VAUT $b - a$.



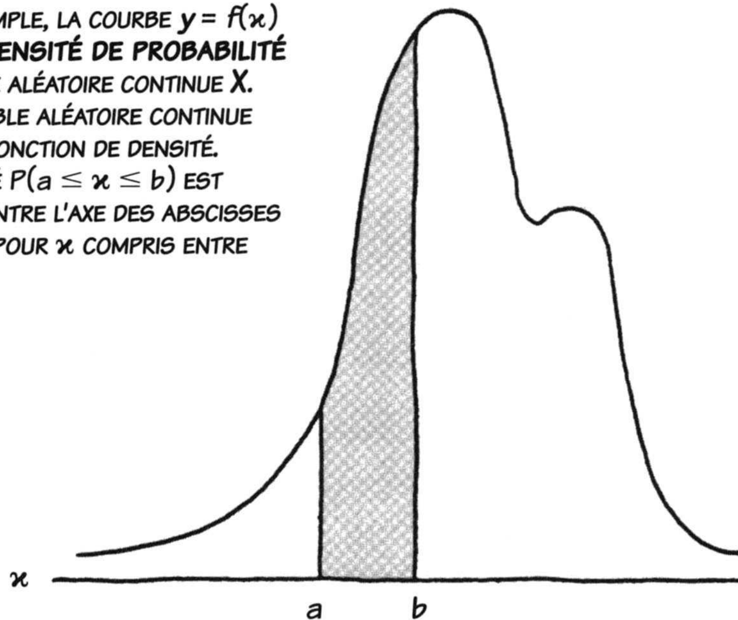
TOUTEFOIS, LA PROBABILITÉ QUE LE POINTEUR TOMBE SUR UNE **VALEUR FIXE** EST DE **ZÉRO** (NOTEZ AUSSI QUE L'AIRE TOTALE SOUS LA COURBE VAUT EXACTEMENT 1).

LES NOMBRES ALÉATOIRES GÉNÉRÉS PAR ORDINATEUR OU AVEC DES CALCULATRICES ONT LE MÊME TYPE DE REPRÉSENTATION. IL SUFFIT DE PRESSER UN BOUTON, ET HOP ! ON OBTIENT UN NOMBRE ENTRE 0 ET 1. CHAQUE NOMBRE EST ÉQUIPROBABLE COMME DANS LE CAS DU POINTEUR DE LA ROUE.



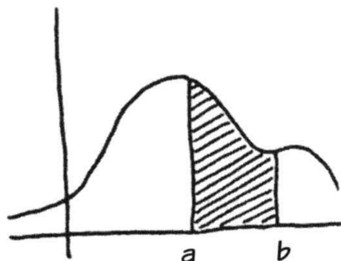
MAIS MALHEUREUSEMENT, CES NOMBRES NE SONT PAS VRAIMENT ALÉATOIRES. ILS SONT PRODUITS PAR DES ALGORITHMES. ON PARLE PLUS PRÉCISÉMENT DE NOMBRES **PSEUDO-ALÉATOIRES.**

DANS CET EXEMPLE, LA COURBE $y = f(x)$ EST APPELÉE **DENSITÉ DE PROBABILITÉ** DE LA VARIABLE ALÉATOIRE CONTINUE X. CHAQUE VARIABLE ALÉATOIRE CONTINUE A SA PROPRE FONCTION DE DENSITÉ. LA PROBABILITÉ $P(a \leq x \leq b)$ EST ALORS L'AIRE ENTRE L'AXE DES ABSCISSES ET LA COURBE POUR x COMPRIS ENTRE a ET b.



EN GÉNÉRAL, LA FONCTION DE DENSITÉ N'EST PAS SI SIMPLE, ET LE CALCUL DE L'AIRE EST LOIN D'ÊTRE TRIVIAL.

$$\int_a^b f(x)\,dx$$



ON UTILISE UNE NOTATION MATHÉMATIQUE POUR DÉCRIRE CETTE AIRE SOUS LA FONCTION $f(x)$. CE SYMBOLE SE LIT « L'INTÉGRALE DE f ENTRE a ET b ».



COMME POUR LES PROBABILITÉS DISCRÈTES, LES FONCTIONS DE DENSITÉS CONTINUES VÉRIFIENT DEUX PROPRIÉTÉS :



$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

(NE VOUS AFFOLEZ PAS DE LA PRÉSENCE DE L'INFINI. ON VEUT JUSTE DIRE QUE L'ON CHERCHE L'AIRE TOTALE SOUS LA COURBE D'UN BOUT À L'AUTRE.)



BIEN QUE LA NOTATION PUISSE SEMBLER ÉTRANGE, ELLE SIGNIFIE SIMPLEMENT UNE AIRE... LE SIGNE D'INTÉGRATION EST LUI-MÊME UN S ALLONGÉ SIGNIFIANT S COMME SOMME.



EN UTILISANT L'INTÉGRALE À LA PLACE DE LA SOMME, ON DÉFINIT LA **MOYENNE et la VARIANCE d'une variable aléatoire continue.**

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx$$

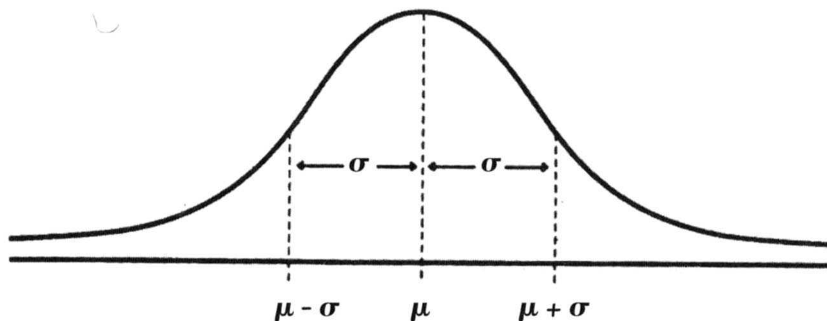
PAR ANALOGIE AVEC LES FORMULES DISCRÈTES :

$$\mu = \sum_{\text{TOUT } x} x p(x)$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx$$

$$\sigma^2 = \sum_{\text{TOUT } x} (x - \mu)^2 p(x)$$

BIEN QUE CELA NE SE VOIE PAS DIRECTEMENT DANS LES FORMULES, LES DÉFINITIONS DE MOYENNE ET DE VARIANCE SONT TOTALEMENT COHÉRENTES AVEC L'INTERPRÉTATION DE TENDANCE CENTRALE ET DE DISPERSION MOYENNE DES PROBABILITÉS DÉFINIES PAR LA DENSITÉ $f(x)$. LE GRAPHIQUE À AVOIR EN TÊTE EST LE SUIVANT :



SOMMATION de variables aléatoires



UNE FOIS CONNUES LA MOYENNE ET LA VARIANCE D'UNE VARIABLE ALÉATOIRE, QUE PEUT-ON EN FAIRE ? TOUT D'ABORD, ON PEUT EN DÉDUIRE LA MOYENNE ET LA VARIANCE D'**AUTRES** VARIABLES ALÉATOIRES.

PAR EXEMPLE, REGARDONS LE LANCER D'UNE PIÈCE. POSONS **X** = 1 SI LA PIÈCE EST FACE ET 0 SI ELLE EST PILE.

x	0	1
$p(x)$	0,5	0,5





À PRÉSENT, VOUS DEVRIEZ POUVOIR CALCULER LA MOYENNE :

$$E[x] = 0 \times p(0) + 1 \times p(1)$$
$$= 0 + 0{,}5$$
$$\mathbf{= 0{,}5}$$

ET LA VARIANCE :

$$\sigma^2 = (0 - 0{,}5)^2\, p(0) + (1 - 0{,}5)^2 p(1)$$
$$\mathbf{= 0{,}25}$$

MAINTENANT JOUONS À UN JEU SIMPLE. D'ABORD VOUS MISEZ 6 € POUR JOUER. JE LANCE UNE PIÈCE, VOUS GAGNEZ 10 € SI C'EST FACE, ET RIEN SI C'EST PILE. AINSI, VOS GAINS G SONT :

$$G = 10X - 6$$



C'EST UNE NOUVELLE VARIABLE ALÉATOIRE ! QUELLES SONT SA MOYENNE ET SA VARIANCE ?

UN PEU DE RÉFLEXION DEVRAIT VOUS CONVAINCRE QUE $E[G]$ VÉRIFIE :

$$E[G] = E[10X - 6]$$
$$= 10E[X] - 6$$

CE QUI REVIENT À :

$$10(0{,}5) - 6 = -1$$

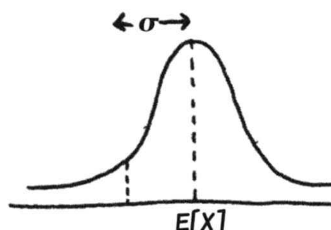
CE QUE L'ON PEUT VÉRIFIER AVEC LE TABLEAU :

x	0	1
G	-6	4
$p(g)$	0,5	0,5



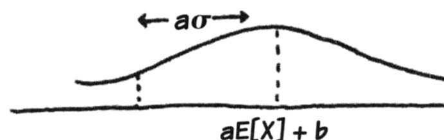
DE FAÇON GÉNÉRALE, IL EST FACILE DE MONTRER QUE :

$$E[aX + b] = aE[X] + b$$



POUR TOUT NOMBRE a ET b ET X, UNE VARIABLE ALÉATOIRE.
POUR LA VARIANCE NOUS AVONS AUSSI LE RÉSULTAT GÉNÉRAL SUIVANT :

$$\sigma^2(aX + b) = a^2\sigma^2(X)$$



DANS LE JEU PRÉCÉDENT, LES RÉSULTATS POSSIBLES SONT -6 ET 4, IL EST DONC ÉVIDENT QUE LA VARIANCE DE G EST PLUS IMPORTANTE QUE CELLE DE X. EN EFFET

$$\sigma^2(G) = \sigma^2(10X + 6)$$
$$= 100\sigma^2(X)$$
$$= 25$$

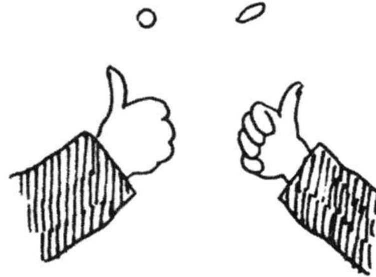
ET

$$\sigma(G) = 5$$



ON PEUT AUSSI AJOUTER DEUX VARIABLES ALÉATOIRES. PAR EXEMPLE, SUPPOSONS QUE L'ON LANCE **DEUX FOIS** UNE PIÈCE. LE NOMBRE DE FACES SUR LES DEUX LANCERS EST $X_1 + X_2$, OÙ X_1 ET X_2 SONT LES VARIABLES ALÉATOIRES DU PREMIER ET DU SECOND LANCER.

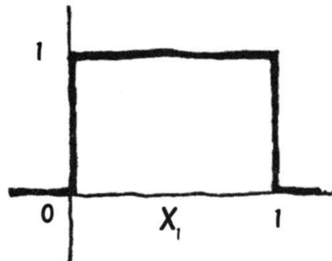
$x_1 + x_2$	0	1	2
$p(x_1 + x_2)$	0,25	0,5	0,25

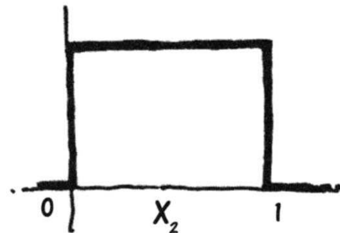


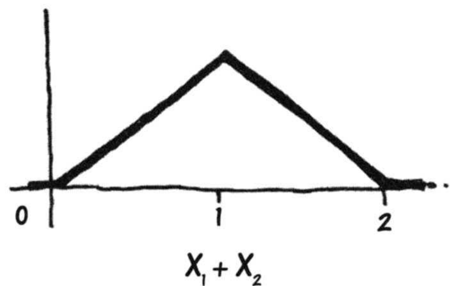
À NOUVEAU, IL EST FACILE DE VOIR QUE :

$$E[X_1 + X_2] = E[X_1] + E[X_2]$$

(NE DEMANDEZ PAS QUELLE EST LA DISTRIBUTION DE PROBABILITÉ DE $X_1 + X_2$, CAR LE LIEN AVEC LES DISTRIBUTIONS ORIGINALES DE X_1 ET X_2 EST COMPLIQUÉ. PAR EXEMPLE, SI X_1 ET X_2 SONT DEUX LANCERS DE POINTEURS SUR UNE ROUE, LA DENSITÉ DE PROBABILITÉ RESSEMBLE À CECI :)







LA VARIANCE DE LA SOMME DE DEUX VARIABLES ALÉATOIRES A UNE FORME SIMPLE LORSQUE LES VARIABLES X ET Y SONT INDÉPENDANTES. LA DÉFINITION TECHNIQUE DE L'**INDÉPENDANCE** EST BASÉE SUR LA PROPRIÉTÉ QUE $P(A \text{ ET } B) = P(A)P(B)$... MAIS, POUR NOUS, L'INDÉPENDANCE SIGNIFIE QUE X ET Y SONT GÉNÉRÉS PAR DES PROCESSUS INDÉPENDANTS, COMME LE LANCER DE PIÈCES OU DE DÉS, ETC.



LORSQUE X ET Y SONT INDÉPENDANTS, LEURS **VARIANCES S'ADDITIONNENT :**

$$\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$$

DANS LE CAS DU LANCER DE DEUX PIÈCES :

$$\sigma^2(X_1 + X_2) = \sigma^2(X_1) + \sigma^2(X_2)$$
$$= 0{,}25 + 0{,}25$$
$$= \mathbf{0{,}50}$$



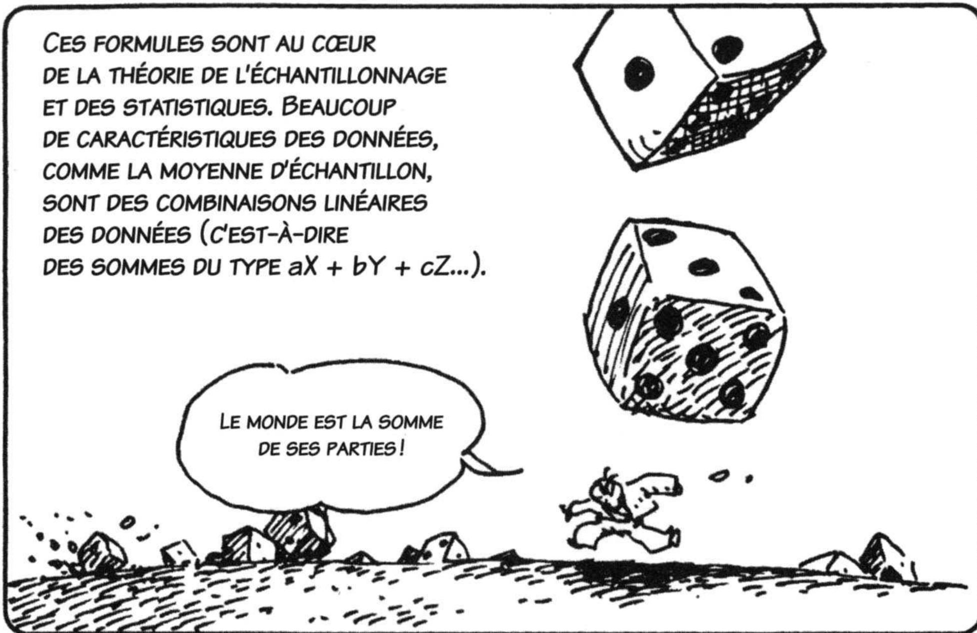
TOUT CELA PEUT SE GÉNÉRALISER À UNE SOMME DE PLUSIEURS VARIABLES ALÉATOIRES :

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

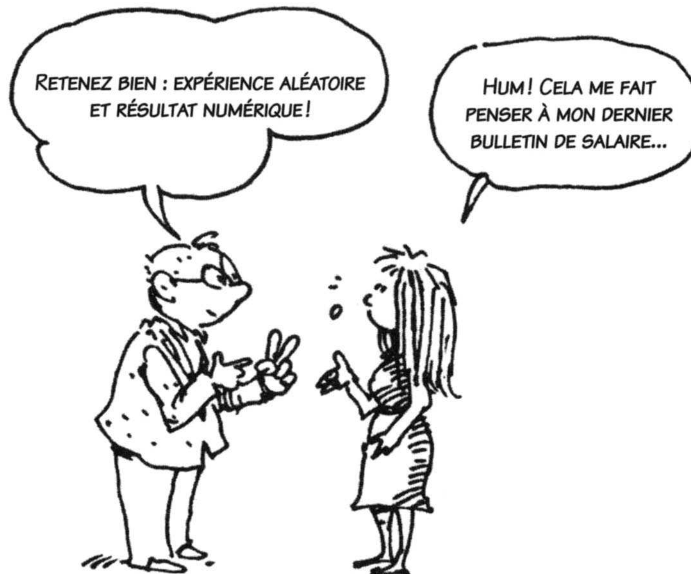
ET LORSQUE LES X_i SONT DES VARIABLES ALÉATOIRES INDÉPENDANTES,

$$\sigma^2\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma^2(X_i)$$



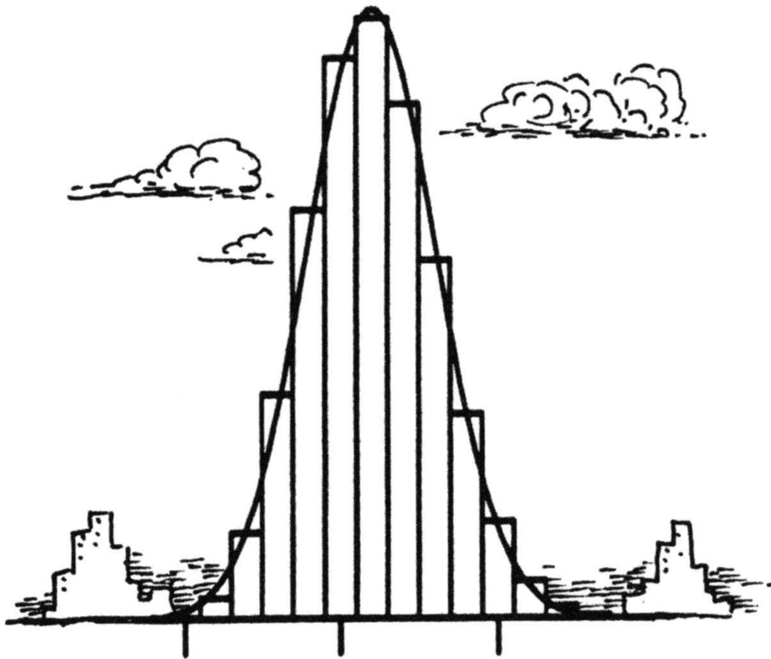


DANS LE PROCHAIN CHAPITRE, NOUS ALLONS VOIR DEUX EXEMPLES IMPORTANTS DE VARIABLES ALÉATOIRES : LE PREMIER, LA **LOI BINOMIALE,** EST UNE SOMME RÉPÉTÉE DE VARIABLES ALÉATOIRES INDÉPENDANTES. LE SECOND, LA **LOI NORMALE,** EST UNE VARIABLE ALÉATOIRE CONTINUE QUI A UN LIEN SURPRENANT AVEC LA LOI BINOMIALE ET AUSSI AVEC D'AUTRES SOMMES DE VARIABLES INDÉPENDANTES.



Chapitre 5
Une histoire de deux distributions

NOUS ALLONS MAINTENANT EXAMINER DEUX EXEMPLES DE VARIABLES ALÉATOIRES, DONT L'UNE EST DISCRÈTE ET L'AUTRE CONTINUE.



Nous allons commencer par la distribution discrète appelée variable aléatoire **binomiale** (ou loi binomiale). Supposons que nous ayons un processus aléatoire avec seulement **deux** résultats : un lancer de pièce, un match sportif, un contrôle de pollution d'automobile. De façon arbitraire, l'un des résultats est appelé **succès** et l'autre **échec.**



C'est ce processus aléatoire aussi appelé épreuve de Bernoulli que nous allons répéter. L'expérience qui consiste à répéter les épreuves est appelée un **schéma de Bernoulli,** si elle vérifie les propriétés suivantes :

1) Le résultat de chaque épreuve est soit un succès, soit un échec.

2) La probabilité p de succès est la même pour chaque épreuve.

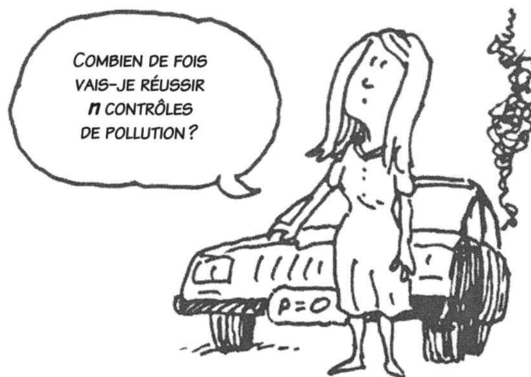
3) Les épreuves sont **indépendantes :** le résultat d'une épreuve n'a aucune influence sur les autres épreuves.



À PARTIR D'UNE ÉPREUVE DE BERNOULLI AYANT UN SUCCÈS AVEC PROBABILITÉ p, ON PEUT CONSTRUIRE UNE NOUVELLE VARIABLE ALÉATOIRE EN RÉPÉTANT LES ÉPREUVES.

La variable aléatoire binomiale X

EST LE **NOMBRE DE SUCCÈS** DANS UN SCHÉMA RÉPÉTÉ DE BERNOULLI À n ÉTAPES, OÙ p EST LA PROBABILITÉ DE SUCCÈS.

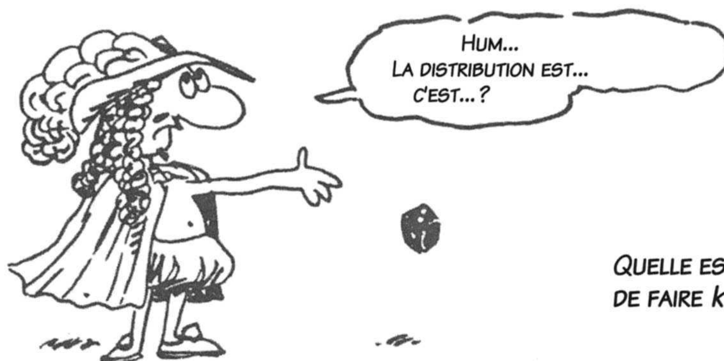


LE NOMBRE DE FACES (SUCCÈS) SUR DEUX LANCERS DE PIÈCE FOURNIT UN EXEMPLE DE VARIABLE ALÉATOIRE BINOMIALE. ICI $n = 2$ et $p = 0{,}5$.

k = NOMBRE DE SUCCÈS	0	1	2
$P(X = k)$	0,25	0,5	0,25



LE PREMIER PARI DE MÉRÉ FOURNIT UN AUTRE EXEMPLE. ON LANCE UN DÉ 4 FOIS DE SUITE. UN SUCCÈS CORRESPOND À FAIRE UN 6.
LA DISTRIBUTION EST...



QUELLE EST LA PROBABILITÉ DE FAIRE k 6 EN 4 LANCERS ?

De façon générale, quelle est la distribution de probabilités d'une variable aléatoire binomiale X de probabilité p **quelconque** et avec n étapes ? Un calcul de probabilité fournit la réponse. La probabilité d'obtenir k succès parmi n essais, $P(X = k)$, est :

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n - k}$$



Ici le symbole $\binom{n}{k}$ se lit « combinaison de k parmi n ». Il est aussi appelé **coefficient binomial.** En France, on le note souvent C_n^k. Il calcule toutes les différentes façons d'obtenir k succès sur n essais. Chaque séquence avec k succès et $n - k$ échecs a une probabilité de $p^k(1 - p)^{n - k}$ par la loi multiplicative. Il y a $\binom{n}{k} = C_n^k$ de ces séquences.

AHHHH...



La formule pour les combinaisons est :

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n - k)!}$$

où

$$n! = 1 \times 2 \times 3 \times ... \times (n - 1) \times (n)$$

Et par convention $0! = 1$. Par exemple, $\binom{4}{2}$ est le nombre de façons de choisir 2 lettres parmi un ensemble de 4. Cela vaut :

$$C_4^2 = \binom{4}{2} = \frac{4!}{2!\,2!} = \frac{24}{4} = 6$$

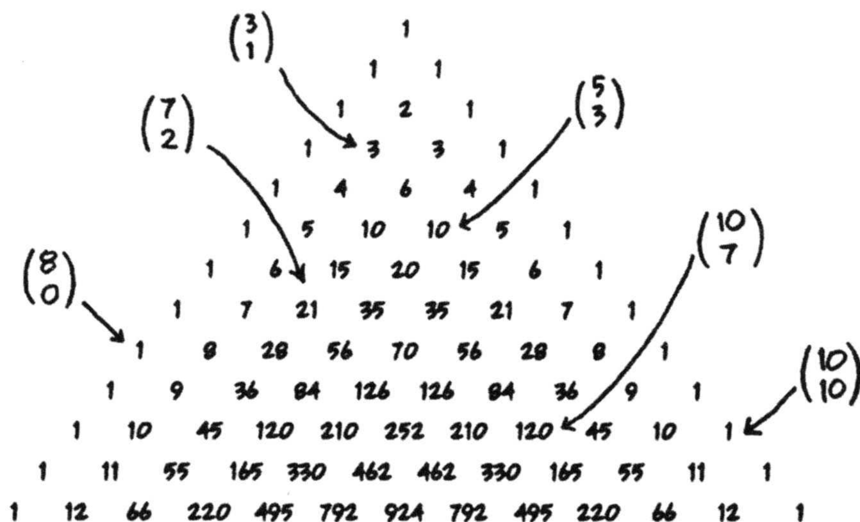
{ABCD}



AB AC AD

BC BD CD

ON PEUT AUSSI UTILISER LE **TRIANGLE DE PASCAL** POUR TROUVER LES COEFFICIENTS BINOMIAUX. CHAQUE ÉLÉMENT EST LA SOMME DES DEUX NOMBRES AU-DESSUS DE LUI.



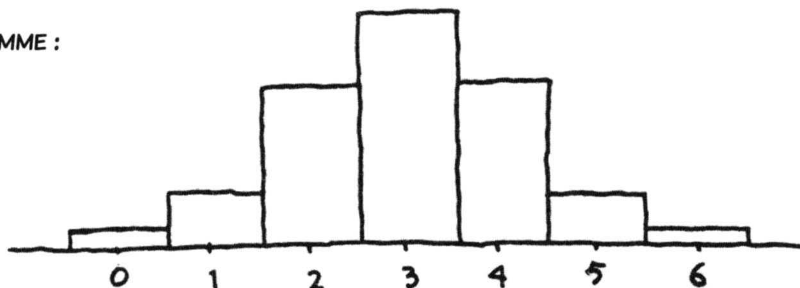
ETC.

POUR TROUVER $\binom{n}{k}$ ALLEZ À LA LIGNE n ET PRENEZ LA k^e VALEUR HORIZONTALEMENT (COMMENCEZ LE COMPTE TOUJOURS PAR ZÉRO).

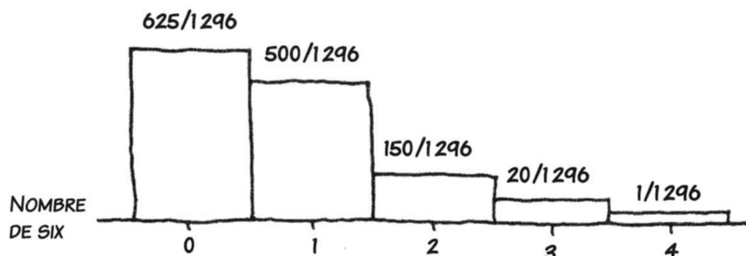
AVEC $p = 0{,}5$, LA DISTRIBUTION DE PROBABILITÉS BINOMIALE EST PARFAITEMENT SYMÉTRIQUE. AINSI, AVEC n, PAR EXEMPLE, ON OBTIENT :

k = NOMBRE DE FACES	0	1	2	3	4	5	6
$P(X = k)$	$\left(\frac{1}{2}\right)^6$	$6\left(\frac{1}{2}\right)^6$	$15\left(\frac{1}{2}\right)^6$	$20\left(\frac{1}{2}\right)^6$	$15\left(\frac{1}{2}\right)^6$	$6\left(\frac{1}{2}\right)^6$	$\left(\frac{1}{2}\right)^6$

D'OÙ CET HISTOGRAMME :



POUR LE LANCER DES 4 DÉS DE MÉRÉ, LA DISTRIBUTION EST PLUS DÉSÉQUILIBRÉE.



LA **MOYENNE** ET LA **VARIANCE** DE LA LOI BINOMIALE SONT :

$$\mu = np$$

$$\sigma^2 = np(1 - p)$$

NOTEZ QUE LA MOYENNE EST INTUITIVE : AVEC n ÉPREUVES DE BERNOULLI, LE NOMBRE ESPÉRÉ DE SUCCÈS SERA np. LA VARIANCE PROVIENT DU FAIT QU'UN SCHÉMA BINOMIAL EST LA SOMME DE n ÉPREUVES DE BERNOULLI INDÉPENDANTES DE VARIANCE $p(1 - p)$.

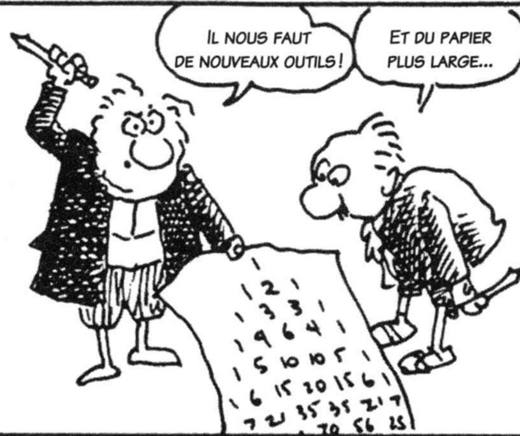


LES PARAMÈTRES D'UNE DISTRIBUTION BINOMIALE SONT n ET p. LA DISTRIBUTION, LA MOYENNE ET LA VARIANCE NE DÉPENDENT QUE DE CES DEUX NOMBRES. ON TROUVE LES TABLES DE LOI BINOMIALE DANS DES MANUELS ET DES PROGRAMMES INFORMATIQUES. VOICI LA TABLE POUR $n = 10$.

VALEUR DE $P(X = k)$

		k										
		0	1	2	3	4	5	6	7	8	9	10
	0,1	0,349	0,387	0,194	0,057	0,011	0,001	0,000	0,000	0,000	0,000	0,000
	0,25	0,056	0,188	0,282	0,250	0,146	0,058	0,016	0,003	0,000	0,000	0,000
p	0,5	0,001	0,010	0,044	0,117	0,205	0,246	0,205	0,117	0,044	0,010	0,001
	0,75	0,000	0,000	0,000	0,003	0,016	0,058	0,146	0,250	0,282	0,188	0,056
	0,9	0,000	0,000	0,000	0,000	0,000	0,001	0,011	0,057	0,194	0,387	0,349

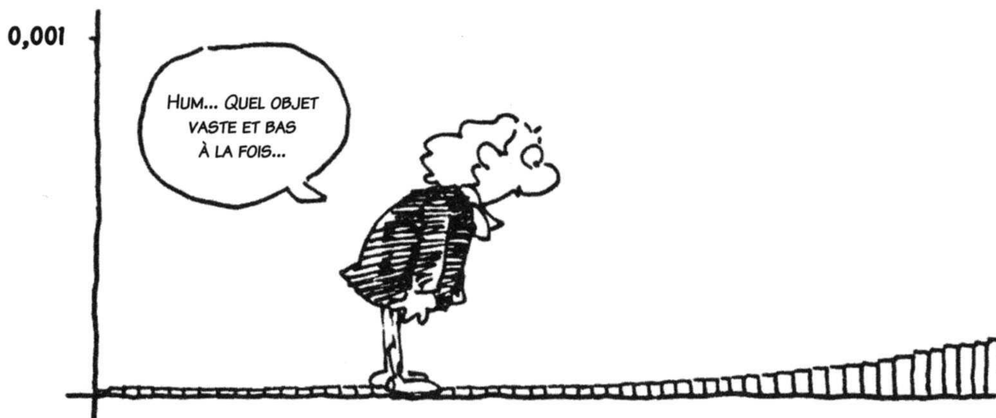
MAIS FAIRE LE CALCUL POUR DE GRANDES VALEURS DE n S'AVÈRE TRÈS DÉLICAT... DU MOINS AU XVIII[e] SIÈCLE, LORSQUE JACQUES BERNOULLI (1654-1705) ET ABRAHAM DE MOIVRE (1667-1754) ESSAYAIENT DE LE FAIRE SANS ORDINATEUR.





MOIVRE DÉPLOYA UNE NOUVELLE ARME DE CALCUL ET MONTRA QUE, LORSQUE $p = \frac{1}{2}$, LA DISTRIBUTION BINOMIALE POUVAIT ÊTRE APPROXIMÉE PAR UNE FONCTION DE DENSITÉ CONTINUE, TRÈS SIMPLE À DÉCRIRE.

POUR VOIR COMMENT CELA FONCTIONNE, IMAGINEZ LA DISTRIBUTION BINOMIALE AVEC $p = 0{,}5$ ET n TRÈS GRAND – DISONS 1 MILLION...



MAINTENANT, DIT MOIVRE, FAITES GLISSER LE GRAPHE AFIN QUE SA MOYENNE SOIT ZÉRO.



ÉCRASEZ LA COURBE LE LONG DE L'AXE DES ABSCISSES AFIN QUE L'ÉCART-TYPE SOIT 1 TOUT EN L'ÉTIRANT SUR L'AXE DES ORDONNÉES POUR QUE L'AIRE SOIT ÉGALE À 1.



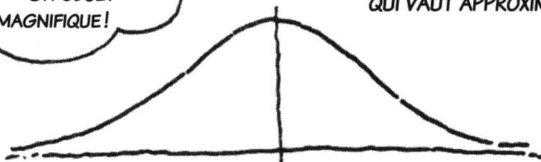
LE RÉSULTAT RESSEMBLE À UNE **COURBE LISSE, SYMÉTRIQUE EN FORME DE CLOCHE,** DONT MOIVRE DONNA L'ÉQUATION SIMPLE :

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

CETTE FONCTION TRÈS IMPORTANTE EST APPELÉE LA **loi normale standard.**

(*e* EST UNE CONSTANTE MATHÉMATIQUE UTILE QUI VAUT APPROXIMATIVEMENT 2,718).





POUR VÉRIFIER QUE LA COURBE EST BIEN EN FORME DE CLOCHE, NOTEZ QUE LOIN DE ZÉRO $f(z)$ EST QUASIMENT NULLE, PAR EXEMPLE $f(-5) = f(5) = 0,0000049$. LA COURBE EST SYMÉTRIQUE, CAR $f(z) = f(-z)$. SON MAXIMUM EST ATTEINT EN ZÉRO OÙ $f(0) = \frac{1}{\sqrt{2\pi}} = 0,39894$.

LA DISTRIBUTION EST APPELÉE LOI NORMALE **STANDARD** CAR LES DÉFORMATIONS ÉTAIENT ORGANISÉES POUR VÉRIFIER DES PROPRIÉTÉS SIMPLES, QUE NOUS PRÉSENTONS SANS PREUVE :

$$\mu = 0$$

$$\sigma = 1$$

POUR RÉSUMER MOIVRE, SI ON « **NORMALISE** » LA DISTRIBUTION BINOMIALE AVEC $p = 1/2$ (AUTREMENT DIT, ON CENTRE EN ZÉRO ET ON FAIT EN SORTE QUE L'ÉCART-TYPE = 1), ALORS ON OBTIENT PRESQUE LA **DISTRIBUTION NORMALE STANDARD.**

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$



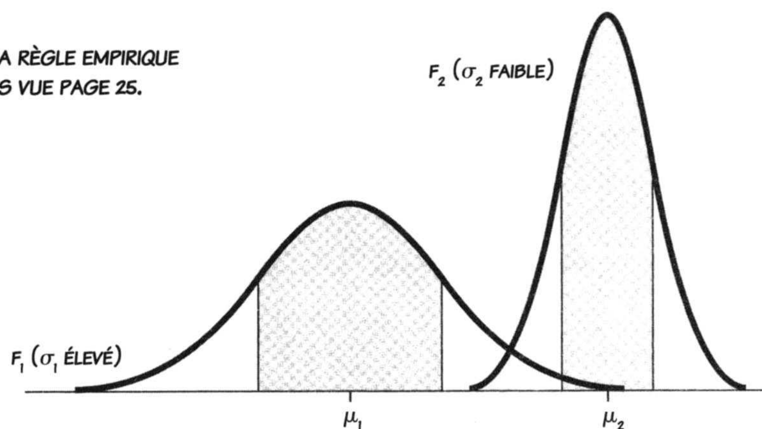
ON PEUT OBTENIR D'AUTRES LOIS NORMALES AVEC DES MOYENNES ET DES VARIANCES DIFFÉRENTES EN ÉTIRANT ET TRANSLATANT LA LOI NORMALE STANDARD. EN GÉNÉRAL, LA FORMULE SUIVANTE NOUS DONNE UNE DISTRIBUTION SYMÉTRIQUE, EN FORME DE CLOCHE CENTRÉE SUR LA MOYENNE μ ET D'ÉCART-TYPE σ :

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

VOICI DEUX DISTRIBUTIONS NORMALES DIFFÉRENTES, LES PARTIES GRISÉES REPRÉSENTENT LA RÉGION SITUÉE À MOINS D'UN ÉCART-TYPE DE LA MOYENNE. LEURS AIRES SONT **ÉGALES.** AINSI, POUR **TOUTE** VARIABLE ALÉATOIRE NORMALE, LA **PROBABILITÉ D'ÊTRE À MOINS D'UN ÉCART-TYPE DE LA MOYENNE EST TOUJOURS LA MÊME,** À SAVOIR À PEU PRÈS 0,68.

$$P(|X - \mu| < \sigma) \approx 0{,}68$$

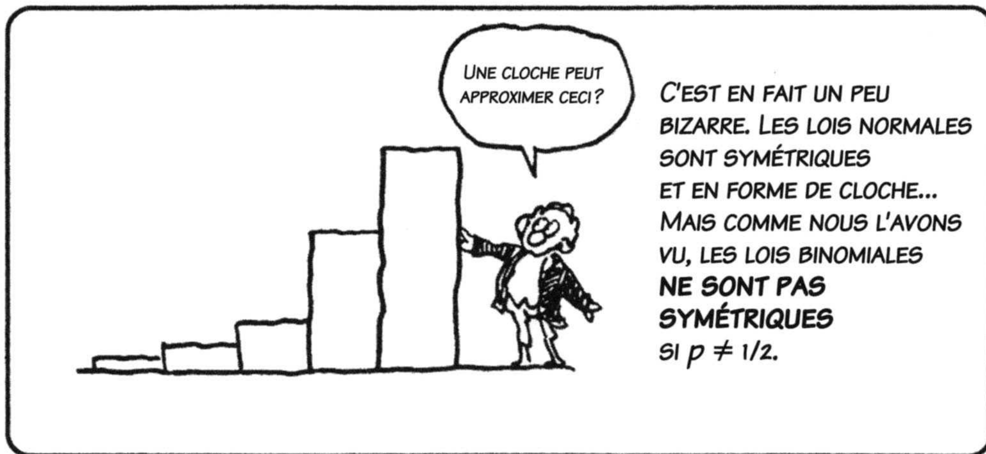
CELA EXPLIQUE LA RÈGLE EMPIRIQUE QUE NOUS AVONS VUE PAGE 25.



MOIVRE PROUVA QUE LA LOI NORMALE STANDARD AVAIT LA FORME DE LA LOI BINOMIALE (NORMALISÉE) POUR $p = 1/2$. MAIS EN FAIT, CELA FONCTIONNE POUR **TOUTE** VALEUR DE p.

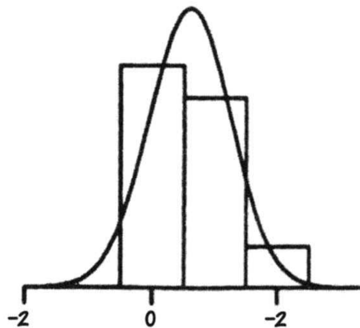
DE FAÇON GÉNÉRALE : QUEL QUE SOIT p, LA DISTRIBUTION BINOMIALE AVEC n ESSAIS DE PROBABILITÉ p EST APPROXIMÉE PAR UNE LOI NORMALE AVEC
$\mu = np$
ET $\sigma = \sqrt{np(1 - p)}$.



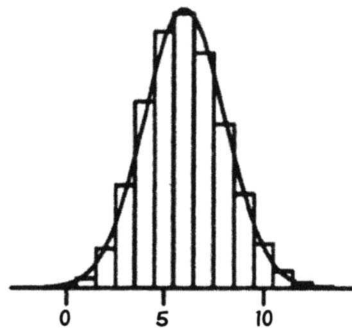


C'EST EN FAIT UN PEU BIZARRE. LES LOIS NORMALES SONT SYMÉTRIQUES ET EN FORME DE CLOCHE... MAIS COMME NOUS L'AVONS VU, LES LOIS BINOMIALES **NE SONT PAS SYMÉTRIQUES** SI $p \neq 1/2$.

MALGRÉ CELA, IL SE TROUVE QUE LORSQUE n AUGMENTE, L'ASYMÉTRIE DE LA LOI BINOMIALE EST CONTRARIÉE. COMME VOUS POUVEZ LE VOIR DANS CET EXEMPLE :



BINOMIALE : $n = 2$ ET $p = 0,3$

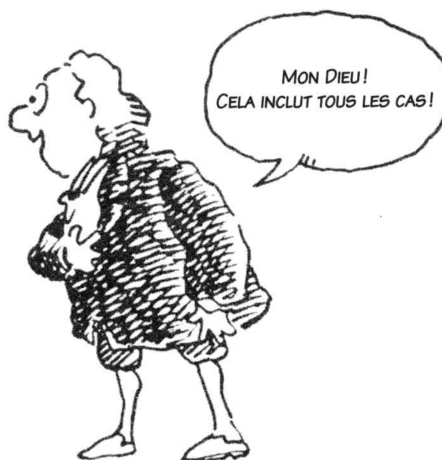


BINOMIALE : $n = 20$ ET $p = 0,3$

EN FAIT, LA DÉCOUVERTE FAITE PAR MOIVRE SUR LA LOI BINOMIALE EST UN CAS PARTICULIER D'UN RÉSULTAT ENCORE PLUS GÉNÉRAL QUI AIDE À MIEUX COMPRENDRE POURQUOI LA LOI NORMALE EST À LA FOIS SI IMPORTANTE ET SI RÉPANDUE DANS LA NATURE.

« Le théorème central limite » :

DES DONNÉES QUI DÉPENDENT DE **NOMBREUX PETITS EFFETS ALÉATOIRES NON CORRÉLÉS** SONT APPROXIMATIVEMENT **DISTRIBUÉES NORMALEMENT.**



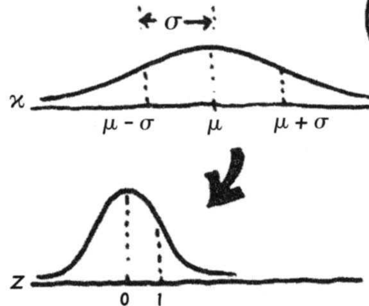
CELA EXPLIQUE QUE LA LOI NORMALE SE RETROUVE **PARTOUT** : DANS LES FLUCTUATIONS DES MARCHÉS FINANCIERS, LE POIDS DES ÉTUDIANTS, LES MOYENNES ANNUELLES DE TEMPÉRATURE, LES RÉSULTATS DU BAC : TOUTES CES QUANTITÉS SONT LE RÉSULTAT D'EFFETS NOMBREUX ET DIFFÉRENTS. PAR EXEMPLE, LE POIDS D'UN ÉTUDIANT EST LE RÉSULTAT DE SES GÈNES, DE SON ALIMENTATION, DE SES MALADIES ET DE SON PASSAGE AU PUB LA VEILLE. QUAND ON AJOUTE TOUS CES EFFETS, ON TROUVE UNE LOI NORMALE ! (RAPPELEZ-VOUS QUE LA LOI BINOMIALE EST LE RÉSULTAT DE n ÉPREUVES INDÉPENDANTES DE BERNOULLI.)



LE Z-SCORE

$$z = \frac{x - \mu}{\sigma}$$

PERMET DE TRANSFORMER UNE LOI NORMALE DE MOYENNE μ ET D'ÉCART-TYPE σ EN UNE LOI NORMALE CENTRÉE RÉDUITE, DE MOYENNE 0 ET D'ÉCART-TYPE 1.



C'EST ENCORE UNE OPÉRATION DE GLISSEMENT ET D'ÉTIREMENT...

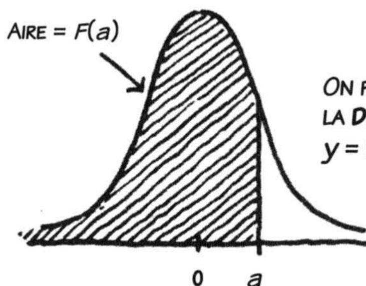


AINSI, POUR TROUVER LA PROBABILITÉ DE N'IMPORTE QUELLE LOI NORMALE, IL SUFFIT DE CONNAÎTRE LA TABLE DE LA LOI NORMALE STANDARD $f(z)$.

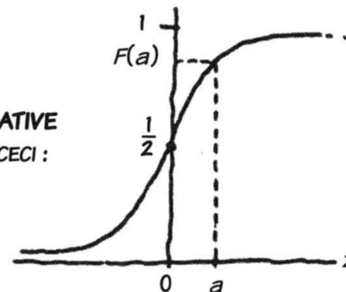
z	−2,5	−2,4	−2,3	−2,2	−2,1	−2,0	−1,9	−1,8	−1,7	−1,6
F(z)	0,006	0,008	0,011	0,014	0,018	0,023	0,029	0,036	0,045	0,055
z	−1,5	−1,4	−1,3	−1,2	−1,1	−1,0	−0,9	−0,8	−0,7	−0,6
F(z)	0,067	0,081	0,097	0,115	0,136	0,159	0,184	0,212	0,242	0,274
z	−0,5	−0,4	−0,3	−0,2	−0,1	0,0	0,1	0,2	0,3	0,4
F(z)	0,309	0,345	0,382	0,421	0,460	0,500	0,540	0,579	0,618	0,655
z	0,5	0,6	0,7	0,8	0,9	1,0	1,1	1,2	1,3	1,4
F(z)	0,691	0,726	0,758	0,788	0,816	0,841	0,864	0,885	0,903	0,919
z	1,5	1,6	1,7	1,8	1,9	2,0	2,1	2,2	2,3	2,4
F(z)	0,933	0,945	0,955	0,964	0,971	0,977	0,982	0,986	0,989	0,992
z	2,5									
F(z)	0,994									



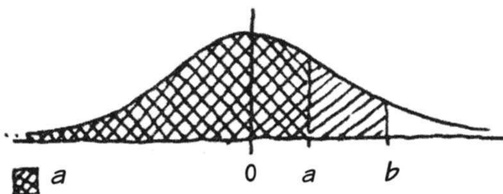
ICI $F(a) = P(z \leq a)$, SOIT L'AIRE COMPRISE ENTRE LA FONCTION DE DENSITÉ ET L'AXE DES ABSCISSES ET À GAUCHE DE $z = a$.



ON PEUT AUSSI TRACER LA **DISTRIBUTION CUMULATIVE** $y = F(z)$ QUI RESSEMBLE À CECI :



LA TABLE DE LOI NOUS PERMET DE CALCULER LA PROBABILITÉ QUE Z SOIT DANS L'INTERVALLE $a \leq z \leq b$. IL S'AGIT DE LA DIFFÉRENCE D'AIRES ENTRE $F(b)$ ET $F(a)$.



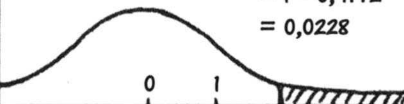
$P(a \leq z \leq b) = F(b) - F(a)$

AINSI, PAR EXEMPLE :

$P(-1 < z < 1) = F(1) - F(-1)$
$= 0{,}8413 - 0{,}1587$
$= 0{,}6826$



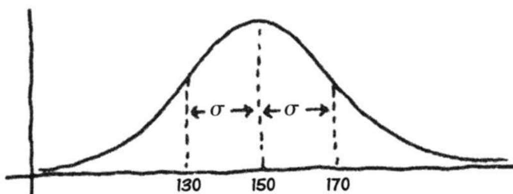
$P(z > 2) = 1 - F(2)$
$= 1 - 0{,}9772$
$= 0{,}0228$



ON PEUT UTILISER CETTE TABLE POUR TROUVER LES PROBABILITÉS DE N'IMPORTE QUELLE DISTRIBUTION NORMALE EN FAISANT LA TRANSFORMATION $z = (x - \mu)/\sigma$.



PAR EXEMPLE, SUPPOSONS QUE LE POIDS DES ÉTUDIANTS SOIT DISTRIBUÉ NORMALEMENT DE MOYENNE $\mu = 150$ ET D'ÉCART-TYPE $\sigma = 20$.

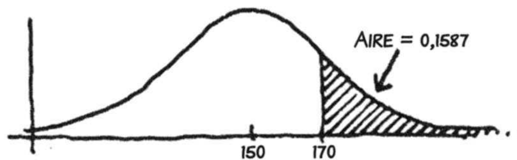


ALORS QUELLE EST LA PROBABILITÉ DE PESER PLUS DE 170 LIVRES ?

MAINTENANT, C'EST « SIMPLEMENT » DE L'ALGÈBRE.

$$P(X > 170)$$
$$= P\left(\frac{x - \mu}{\sigma} > \frac{170 - 150}{20}\right)$$
$$= P\left(z > \frac{20}{20}\right)$$
$$= P(z > 1)$$

IL S'AGIT DE $1 - F(1)$. EN LISANT LA TABLE ON OBTIENT $1 - 0{,}8413 = 0{,}1587$:

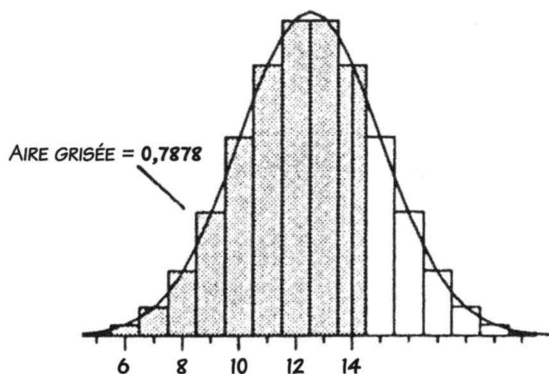


UN PEU MOINS D'UN ÉTUDIANT SUR SIX FAIT PENCHER LA BALANCE AU-DESSUS DE 170 LIVRES (SOIT 77 kg).

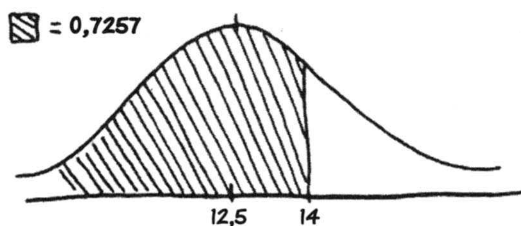
LA RÈGLE GÉNÉRALE POUR CALCULER DES PROBABILITÉS NORMALES EST DONC :

$$P(a \leq X \leq b) = F\left(\frac{b - \mu}{\sigma}\right) - F\left(\frac{a - \mu}{\sigma}\right)$$

MAINTENANT REVENONS À L'APPROXIMATION BINOMIALE DE MOIVRE. REGARDONS LA DISTRIBUTION BINOMIALE AVEC $p = \frac{1}{2}$ ET $n = 25$ (DISONS 25 TIRAGES DE PIÈCES). ON PEUT CALCULER (OU LIRE DANS UNE TABLE) N'IMPORTE QUELLE PROBABILITÉ. PAR EXEMPLE $P(X \leq 14)$ EST EXACTEMENT ÉGAL À **0,7878.**



CALCULONS MAINTENANT UNE VARIABLE ALÉATOIRE NORMALE X^* AVEC LA MÊME MOYENNE $\mu = np = (25)(0,5) = 12,5$ ET UN ÉCART-TYPE $\sigma = \sqrt{np(1-p)} = 2,5$.



$$P(X^* \leq 14) = P\left(Z \leq \frac{14 - 12,5}{2,5}\right)$$

$$= P(Z \leq 0,6)$$

$$\mathbf{= 0,7257}$$



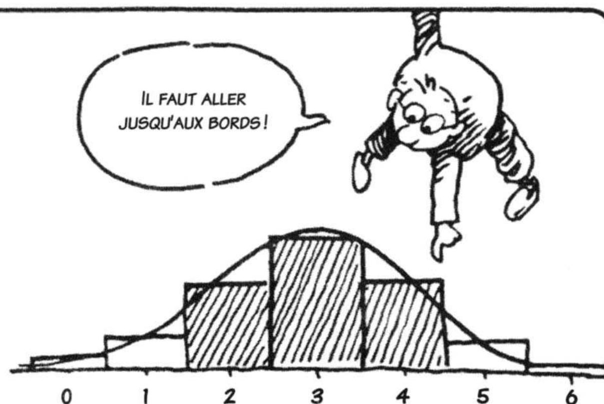
AH, MAIS ON PEUT FAIRE MIEUX ! SI ON REGARDE ATTENTIVEMENT LE PREMIER HISTOGRAMME, ON VOIT QUE LES BARRES SONT CENTRÉES SUR LES ENTIERS. CELA VEUT DIRE QUE $P(X^* \leq 14)$ EST EN FAIT L'AIRE SITUÉE SOUS LES BARRES EN DESSOUS DE x = **14,5.** NOUS DEVONS DONC INCLURE CE 0,5. AINSI,

$$P(X^* \leq 14,5) = P(z \leq 0,8)$$

$$\mathbf{= 0,7881}$$

UNE TRÈS BONNE APPROXIMATION DE 0,7878, EN EFFET !

CE 0,5 SUPPLÉMENTAIRE S'APPELLE LA **correction de continuité.** NOUS DEVONS L'INCLURE POUR OBTENIR UNE BONNE APPROXIMATION CONTINUE DE NOTRE LOI BINOMIALE DISCRÈTE X. CELA SE RÉSUME AVEC UNE ÉQUATION UN PEU HORRIBLE MAIS SIMPLE D'UTILISATION.

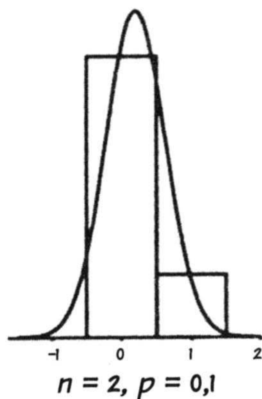


$$P(a \leq X \leq b) \approx P\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

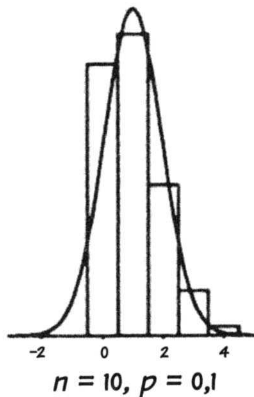
QUAND CETTE APPROXIMATION EST-ELLE « SUFFISAMMENT BONNE » ? POUR LES STATISTICIENS LA RÈGLE EMPIRIQUE VEUT QUE n SOIT SUFFISAMMENT GRAND POUR QUE LES NOMBRES ESPÉRÉS DES SUCCÈS ET DES ÉCHECS SOIENT TOUS DEUX **SUPÉRIEURS À 5 :**

$$np \geq 5 \qquad \text{ET} \qquad n(1-p) \geq 5$$

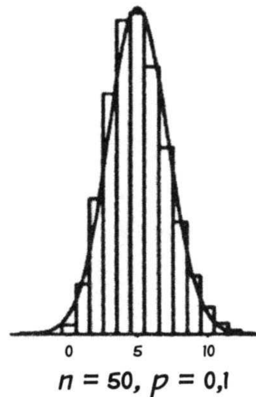
LES HISTOGRAMMES SUIVANTS ILLUSTRENT QUE, LORSQUE $p = 0{,}1$, L'APPROXIMATION EST MÉDIOCRE (VOIRE PIRE) JUSQU'À CE QUE n ATTEIGNE 50, POUR LEQUEL $np = 5$.



$n = 2,\ p = 0{,}1$



$n = 10,\ p = 0{,}1$



$n = 50,\ p = 0{,}1$

POURQUOI L'APPROXIMATION BINOMIALE EST-ELLE SI IMPORTANTE ? EN FAIT LA DISTRIBUTION BINOMIALE APPARAÎT SOUVENT DANS LA NATURE, CE QUI N'EST PAS SI COMPLIQUÉ À COMPRENDRE. MAIS ELLE PEUT ÊTRE FASTIDIEUSE À CALCULER.



IL Y EN A UNE NOUVELLE POUR CHAQUE VALEUR DE n ET DE p...

LA LOI NORMALE QUI L'APPROXIME EST PEUT-ÊTRE MOINS INTUITIVE, MAIS ELLE EST TRÈS SIMPLE D'UTILISATION. GRÂCE À LA TRANSFORMATION EN z, ON PEUT CONVERTIR **TOUTES** LES LOIS NORMALES EN LOI NORMALE STANDARD, ET ON PEUT DONC TROUVER LES PROBABILITÉS ASSOCIÉES DIRECTEMENT À PARTIR D'UNE SEULE TABLE NUMÉRIQUE.

DANS UN LIVRE OU SUR L'ÉCRAN D'UN ORDINATEUR !



EN OUTRE, LA LOI NORMALE EST VRAIMENT LA **MÈRE DE TOUTES LES DISTRIBUTIONS** !

C'EST LE THÉORÈME CENTRAL LIMITE !

MAMAN !
MAMAN !

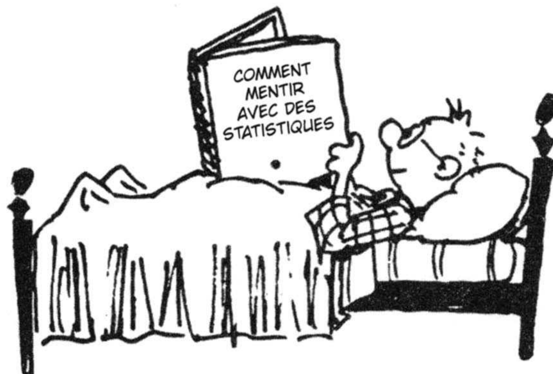


Chapitre 6
Échantillonnage

À PRÉSENT, APRÈS UN RÉGIME VARIÉ DE PIÈCES, DE DÉS ET D'IDÉES ABSTRAITES VOUS DEVEZ VOUS DEMANDER EN QUOI LES OUTILS STATISTIQUES QUE NOUS AVONS CONSTRUITS PEUVENT NOUS AIDER DANS LE **MONDE RÉEL.** EH BIEN, NOUS ALLONS ENFIN LE DÉCOUVRIR...

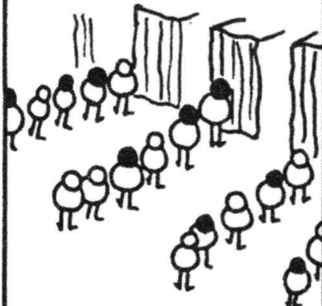


DANS CE CHAPITRE, NOUS COMMENÇONS À REGARDER LE **VRAI** CŒUR DU BUSINESS DES STATISTIQUES, DONT LE BUT, APRÈS TOUT, EST DE FAIRE GAGNER DU **TEMPS** ET DE L'**ARGENT**. LES GENS DÉTESTENT PERDRE LEUR TEMPS EN FAISANT UN **TRAVAIL INUTILE.** S'IL Y A UNE CHOSE DONT LES STATISTIQUES SONT CAPABLES, C'EST JUSTEMENT DE NOUS INDIQUER JUSQU'À QUEL POINT NOUS POUVONS NOUS PERMETTRE D'ÊTRE PARESSEUX.

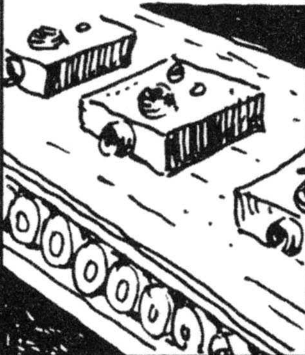


Le problème avec notre monde est que les collections de choses sont tellement vastes qu'il est difficile d'obtenir l'information dont on a besoin.

Population des votants : quel pourcentage pour quel candidat ?



Produits manufacturés : quelle proportion de produits défectueux ?



Cornichons : quelle est leur longueur moyenne ?



L'information est utile aux fabricants de bocaux de cornichons !

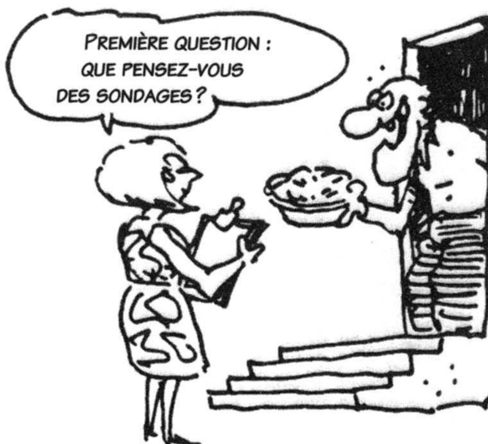
La réponse d'un castor appliqué, travailleur et candide serait de mesurer chaque cornichon du monde et de faire un peu d'arithmétique.



Mais nous ne sommes pas des castors – nous sommes des statisticiens ! Nous cherchons un moyen plus facile.



Une méthode est de prendre un **échantillon...** Un sous-ensemble relativement petit de la population totale, à la façon de ce que font les sondeurs pour les élections.



Une question évidente apparaît : quelle doit être la taille de l'échantillon pour avoir des résultats significatifs ?



Et la réponse à cette question, que vous devriez graver dans votre cerveau, est : si n est le nombre d'éléments de l'échantillon, alors tout dépend de la valeur :

$$\frac{1}{\sqrt{n}}$$

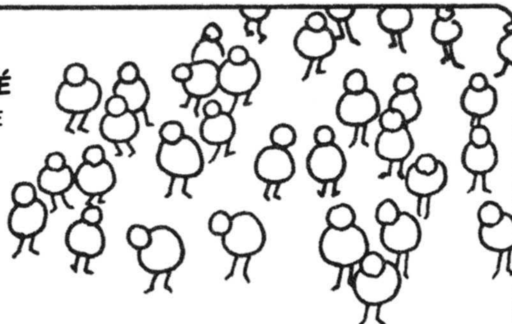


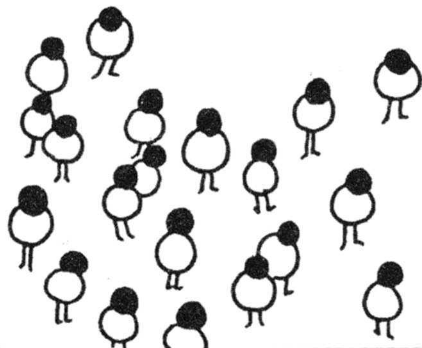
MÉTHODE D'ÉCHANTILLONNAGE



AVANT D'EN VENIR AUX NOMBRES, NOUS DEVONS SIGNALER QUE LA **QUALITÉ** DE L'ÉCHANTILLON EST AUSSI IMPORTANTE QUE SA **TAILLE.**

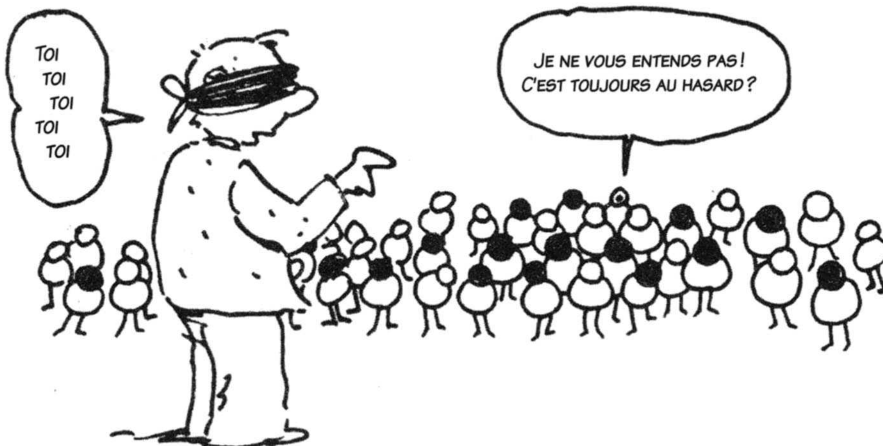
COMMENT S'ASSURER QUE NOUS CHOISISSONS UN ÉCHANTILLON REPRÉSENTATIF ?





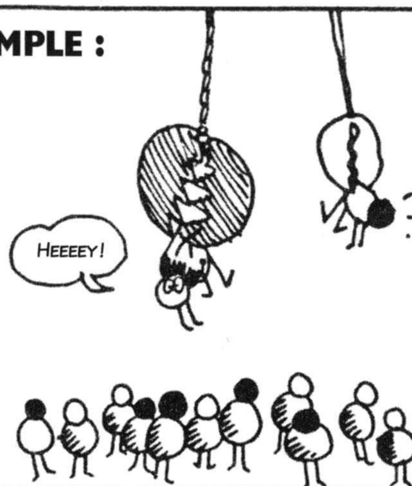
LE PROCESSUS DE **SÉLECTION** EST **LUI-MÊME** ESSENTIEL. PAR EXEMPLE, UN SONDAGE ÉLECTORAL QUI EXCLURAIT SYSTÉMATIQUEMENT LES PARISIENS SERAIT INUTILE. IL Y A BIEN D'AUTRES MANIÈRES DE GÂCHER OU DE BIAISER UN ÉCHANTILLON.

NE PROLONGEONS PAS LE MYSTÈRE : LA FAÇON D'OBTENIR DES RÉSULTATS STATISTIQUES FIABLES EST DE PRENDRE UN ÉCHANTILLON AU **hasard.**



Échantillonnage aléatoire SIMPLE :

IL NÉCESSITE UNE GRANDE POPULATION D'ÉLÉMENTS ET UNE PROCÉDURE POUR EN CHOISIR n PARMI EUX. SI LA PROCÉDURE ASSURE QUE **CHACUN DES ÉCHANTILLONS POSSIBLES** DE n ÉLÉMENTS EST **ÉQUIPROBABLE,** ALORS ELLE NOUS PROCURE UN **échantillon aléatoire simple.**



L'ÉCHANTILLONNAGE ALÉATOIRE SIMPLE DÉFINIT UN STANDARD PAR RAPPORT AUX AUTRES MÉTHODES, CAR IL VÉRIFIE DEUX PROPRIÉTÉS :



1) NON BIAISÉ : CHAQUE ÉLÉMENT A LES MÊMES CHANCES D'ÊTRE SÉLECTIONNÉ.

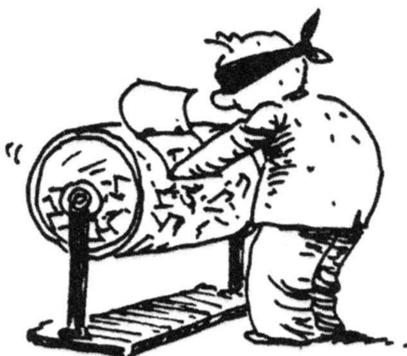
2) INDÉPENDANCE : LA SÉLECTION D'UN ÉLÉMENT N'A AUCUNE INCIDENCE SUR LA SÉLECTION DES AUTRES ÉLÉMENTS.

MALHEUREUSEMENT, IL EST DIFFICILE DE TROUVER DES ÉCHANTILLONS INDÉPENDANTS ET TOTALEMENT SANS BIAIS DANS LE MONDE RÉEL. PAR EXEMPLE, UN SONDAGE EFFECTUÉ EN COMPOSANT AU HASARD DES NUMÉROS DE TÉLÉPHONE SERA BIAISÉ : IL IGNORE LES PERSONNES NE DÉTENANT PAS DE TÉLÉPHONE ET SURREPRÉSENTE CELLES AYANT PLUS D'UNE LIGNE TÉLÉPHONIQUE.



THÉORIQUEMENT, IL EST POSSIBLE D'OBTENIR UN ÉCHANTILLON ALÉATOIRE SIMPLE EN ÉTABLISSANT LA **BASE D'ÉCHANTILLONNAGE :** UNE LISTE DE TOUS LES ÉLÉMENTS DE LA POPULATION. ON PEUT ALORS AVEC UN GÉNÉRATEUR DE NOMBRES ALÉATOIRES SÉLECTIONNER n ÉLÉMENTS AU HASARD.





DE MÊME, ON PEUT INSCRIRE LES NOMS SUR DES CARTES ET EN PRÉLEVER n AU HASARD DANS UNE URNE OU UN TAMBOUR.

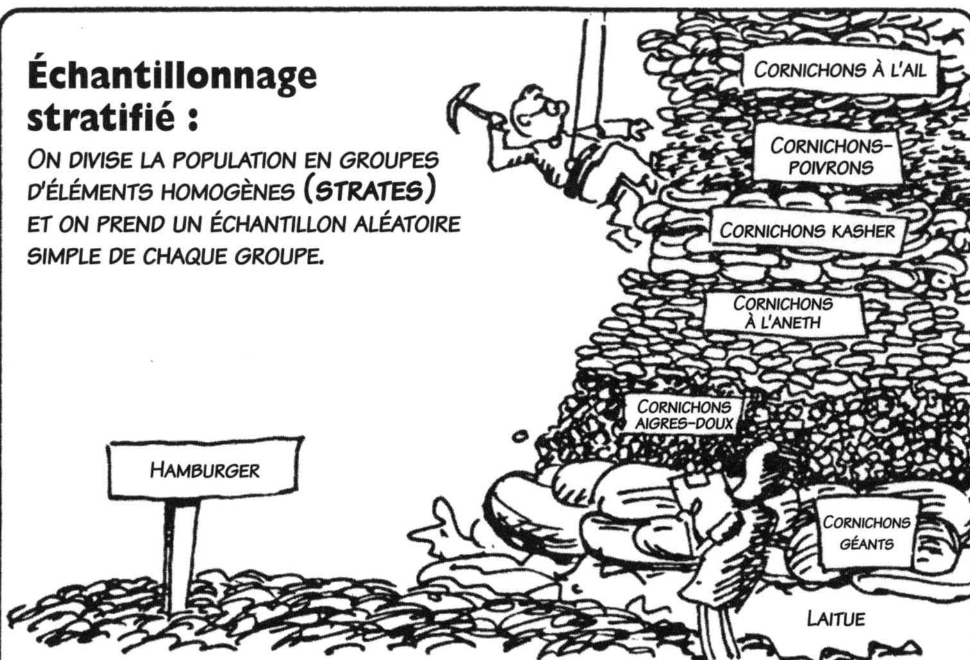
MAIS CE N'EST PAS TOUJOURS FACILE. IL PEUT ÊTRE TROP COÛTEUX, POLÉMIQUE OU MÊME IMPOSSIBLE DE CRÉER LA BASE D'ÉCHANTILLONNAGE. PAR EXEMPLE, UNE ÉTUDE MINISTÉRIELLE SUR LA QUALITÉ DES EAUX NÉCESSITE UNE BASE D'ÉCHANTILLONNAGE DES LACS. IL FAUT DONC QUE QUELQU'UN DÉCIDE :



Y A-T-IL D'AUTRES FAÇONS PLUS EFFICACES ET MOINS COÛTEUSES DE CRÉER UN ÉCHANTILLON ? LA RÉPONSE EST OUI SI VOUS CONNAISSEZ UN PEU VOTRE POPULATION. PAR EXEMPLE...

Échantillonnage stratifié :

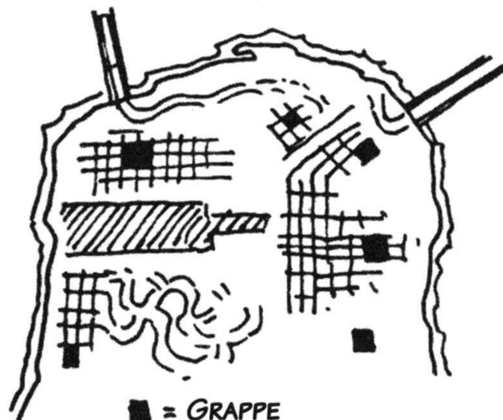
ON DIVISE LA POPULATION EN GROUPES D'ÉLÉMENTS HOMOGÈNES (**STRATES**) ET ON PREND UN ÉCHANTILLON ALÉATOIRE SIMPLE DE CHAQUE GROUPE.



PAR EXEMPLE, LA POPULATION DE TOUS LES **CORNICHONS** PEUT ÊTRE STRATIFIÉE PAR **TYPES.** LEUR TAILLE SERA ALORS MOINS VARIABLE À L'INTÉRIEUR DE CHAQUE STRATE.

Échantillonnage en grappes :

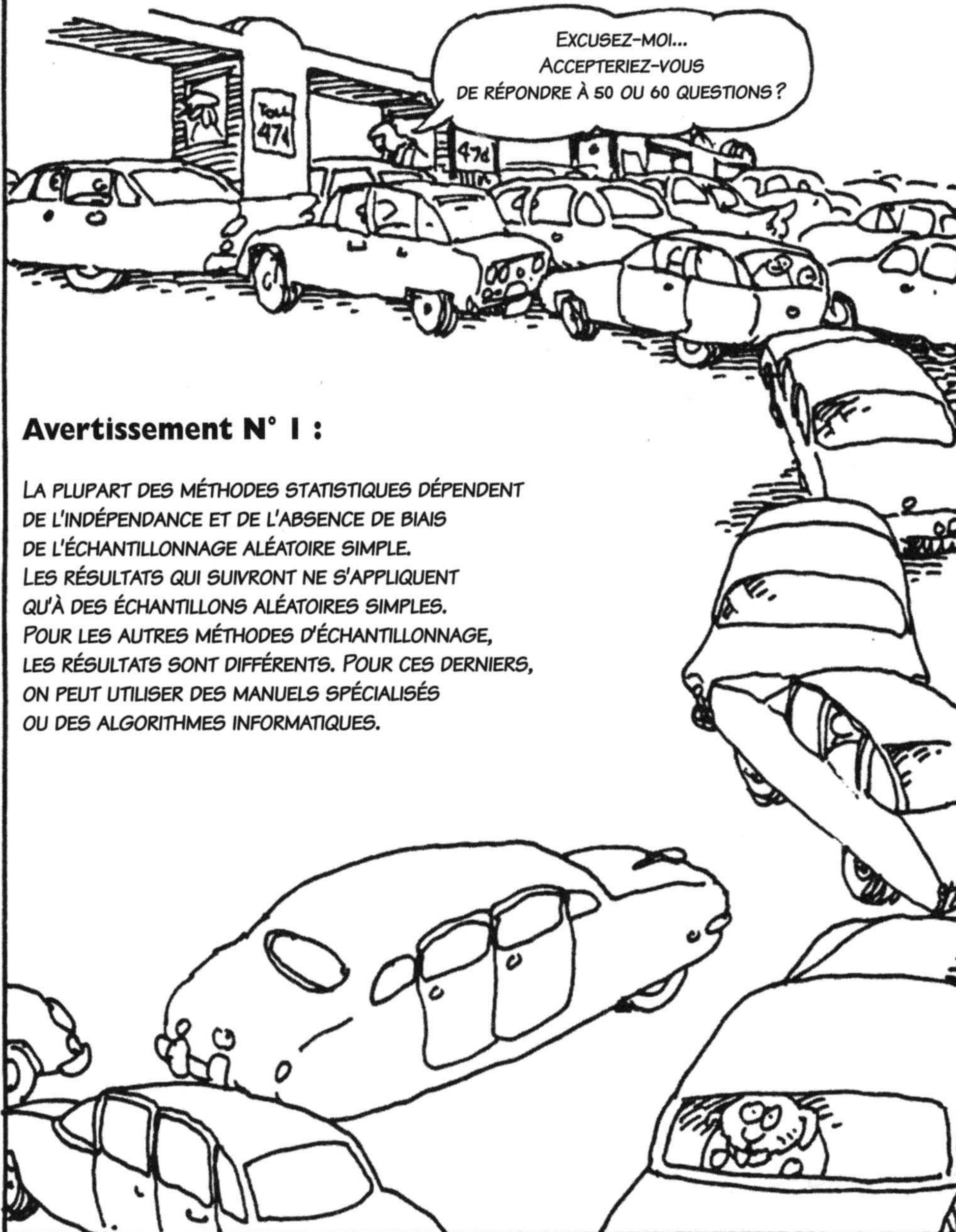
ON SUBDIVISE LA POPULATION EN PLUS PETITES GRAPPES. ON PREND ALORS UN ÉCHANTILLON ALÉATOIRE SIMPLE DE GRAPPES ET ON OBSERVE TOUS LES ÉLÉMENTS DES GRAPPES SÉLECTIONNÉES. CETTE TECHNIQUE EST TRÈS RENTABLE LORSQUE LES COÛTS DE TRANSPORT ENTRE LES ÉLÉMENTS SONT ÉLEVÉS.



POUR UNE ÉTUDE SUR LES FOYERS D'UNE VILLE, ON PEUT PAR EXEMPLE : DIVISER LA VILLE EN BLOCS, SÉLECTIONNER UN ÉCHANTILLON DE BLOCS ET SONDER CHAQUE FOYER DES BLOCS CHOISIS.

Échantillonnage systématique :

On commence par choisir un premier élément au hasard, puis on prend tous les k^e suivants. Par exemple, une **étude sur le trafic autoroutier** peut tester chaque voiture sur 100 passages à une barrière de péage. Cette stratégie est facile à mettre en place et elle est plus efficace si la circulation varie peu avec le temps.



Avertissement N° 1 :

La plupart des méthodes statistiques dépendent de l'indépendance et de l'absence de biais de l'échantillonnage aléatoire simple. Les résultats qui suivront ne s'appliquent qu'à des échantillons aléatoires simples. Pour les autres méthodes d'échantillonnage, les résultats sont différents. Pour ces derniers, on peut utiliser des manuels spécialisés ou des algorithmes informatiques.

Avertissement N° 2 :





SANS CRÉATION DE HASARD, IL N'Y A PAS D'ANALYSE STATISTIQUE FIABLE, ET CE QUELLES QUE SOIENT LES ADAPTATIONS. L'AVANTAGE DE L'ÉCHANTILLONNAGE ALÉATOIRE EST QUE CELUI-CI GARANTIT STATISTIQUEMENT LA RIGUEUR DES ENQUÊTES.

UNE MÉTHODE TRÈS SOUVENT UTILISÉE ET PARTICULIÈREMENT SUJETTE À BIAIS EST CELLE DE L'**échantillonnage de commodité.** L'ENQUÊTEUR ÉVITE LES TRACAS D'UNE PROCÉDURE PARTICULIÈRE EN PRENANT SIMPLEMENT LES *n* PREMIERS ÉLÉMENTS DE LA POPULATION QUI LUI TOMBENT SOUS LA MAIN.

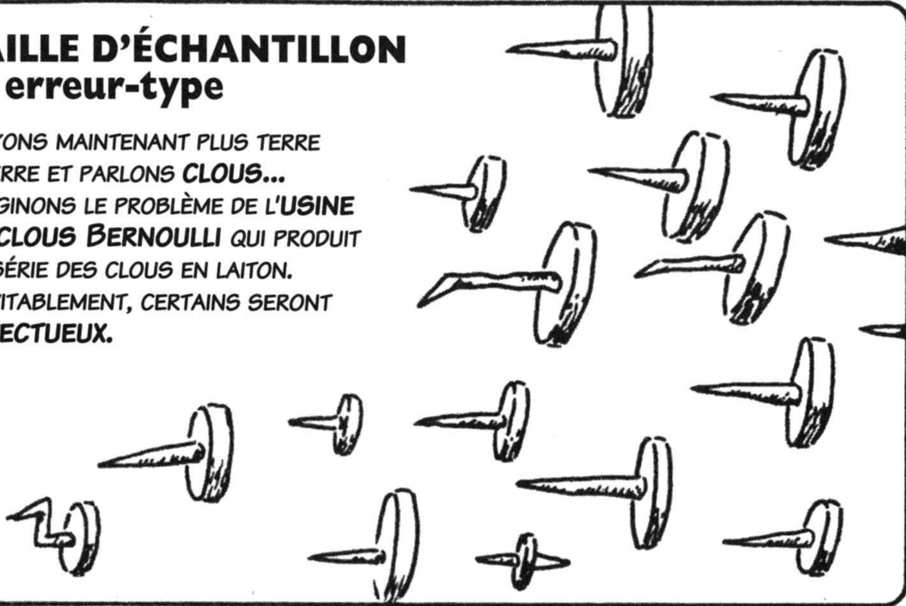


ON EN TROUVE UN EXEMPLE DANS LE LIVRE ***LES FEMMES ET L'AMOUR*** ÉCRIT EN 1987 PAR SHERE HITE. EN EFFET, 100 000 QUESTIONNAIRES FURENT ENVOYÉS À DES ORGANISATIONS FÉMININES (UN **ÉCHANTILLON D'OPPORTUNITÉ**). SEULEMENT **4,5 %** FURENT RENVOYÉS (**BIAIS DANS LES RÉPONSES**). SES « RÉSULTATS » ÉTAIENT DONC FONDÉS SUR UN ÉCHANTILLON DE FEMMES MOTIVÉES, DÉSIREUSES DE RÉPONDRE AUX QUESTIONS DE L'ENQUÊTE POUR DES RAISONS DIVERSES.

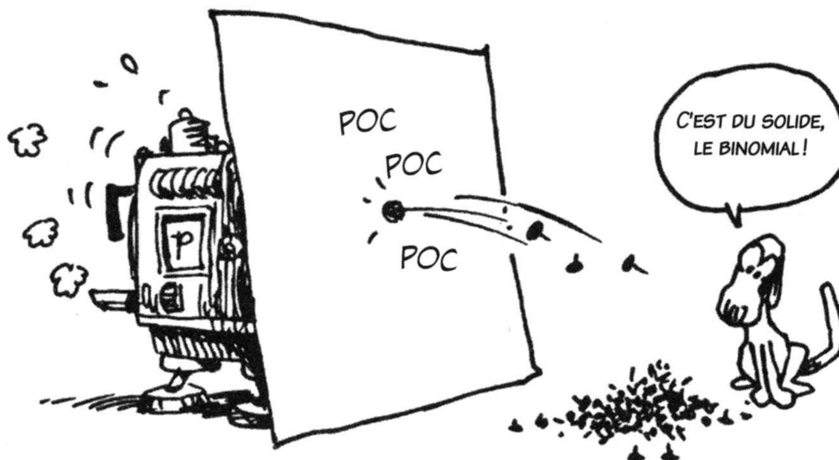


TAILLE D'ÉCHANTILLON et erreur-type

SOYONS MAINTENANT PLUS TERRE À TERRE ET PARLONS **CLOUS...** IMAGINONS LE PROBLÈME DE L'**USINE DE CLOUS BERNOULLI** QUI PRODUIT EN SÉRIE DES CLOUS EN LAITON. INÉVITABLEMENT, CERTAINS SERONT **DÉFECTUEUX.**



LES LECTEURS ATTENTIFS RECONNAÎTRONT UN **SCHÉMA DE BERNOULLI :** CHAQUE NOUVEAU CLOU EST LE RÉSULTAT D'UNE ÉPREUVE DE BERNOULLI AVEC UNE PROBABILITÉ p DE SUCCÈS (CLOU SANS DÉFAUT) ET UNE PROBABILITÉ $1 - p$ D'ÉCHEC (CLOU DÉFECTUEUX).



ON PEUT CONSIDÉRER QUE TOUT SE PASSE COMME SI UNE **MACHINE DE BERNOULLI CACHÉE, MAIS RÉELLE,** PRODUISAIT SELON UNE PROBABILITÉ p LES RÉSULTATS OBSERVÉS DANS LE MONDE RÉEL.

COMME LA MACHINE DE BERNOULLI EST INVISIBLE, ON NE CONNAÎT PAS p MAIS ON AIMERAIT LE CONNAÎTRE. ON PREND DONC UN **ÉCHANTILLON ALÉATOIRE** DE n CLOUS, ET ON REMARQUE QUE x D'ENTRE EUX SONT IRRÉPROCHABLES.



MAINTENANT LA PROPORTION DE SUCCÈS DANS L'ÉCHANTILLON DEVRAIT ÊTRE DE L'ORDRE DE p. ON NOTE CETTE PROPORTION $\hat{p}$ ET ON PRONONCE « p-CHAPEAU ».

$$\hat{p} = \frac{x}{n}$$

$\hat{p}$ EST LE RATIO DU NOMBRE x DE SUCCÈS DE L'ÉCHANTILLON PAR LA TAILLE n DE L'ÉCHANTILLON. PAR EXEMPLE, SI p VAUT 0,85 ET QUE L'ON AIT ÉCHANTILLONNÉ n = 1000 CLOUS, ON POURRAIT TROUVER x = 832 CLOUS SANS DÉFAUT, DE SORTE QUE $\hat{p}$ = 0,832.

LA QUESTION EST : CETTE ESTIMATION EST-ELLE BONNE ?



ET NOUS ALLONS RÉPONDRE À UNE AUTRE QUESTION : QUE SIGNIFIE NOTRE PREMIÈRE QUESTION ?

On ne peut pas connaître la véritable différence entre $\hat{p}$ et p, car nous ne savons pas la valeur exacte de p. La **vraie** question est : si l'on prend **plusieurs échantillons de 1000 clous** et que l'on observe chaque $\hat{p}$, comment ces $\hat{p}$ seront-ils distribués autour de p?



En fait, ces valeurs de $\hat{p}$ ressemblent fortement à une **variable aléatoire :** la sélection d'échantillon de n éléments est une expérience aléatoire et l'observation de $\hat{p}$ en est le résultat numérique !



Pour être précis, si X est le nombre de succès de l'échantillon, alors X n'est rien d'autre qu'une variable aléatoire binomiale (avec n essais et de probabilité p)... et nous définissons la **proportion d'échantillon** (ou proportion observée) comme étant la variable aléatoire :

$$\hat{P} = \frac{X}{n}$$



CONNAISSANT X, ON CONCLUT RAPIDEMENT QUELQUES PROPRIÉTÉS SUR $\hat{P}$:

1) LA MOYENNE DE $\hat{P}$ EST $E[\hat{P}] = p$
2) L'ÉCART-TYPE DE $\hat{P}$ EST

$$\sigma(\hat{P}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

3) POUR DE GRANDES VALEURS DE n, $\hat{P}$ EST APPROXIMATIVEMENT NORMALE



ET VOILÀ VOUS SAVEZ TOUT ! LES VALEURS OBSERVÉES DE $\hat{P}$ SONT CENTRÉES EN p (SANS SURPRISE), ET L'ÉCART-TYPE (OU EN ANGLAIS SPREAD) EST PROPORTIONNEL À CE NOMBRE MAGIQUE QUE NOUS AVONS MENTIONNÉ EN DÉBUT DE CHAPITRE.



$$\frac{1}{\sqrt{n}}$$



ET COMME $\hat{P}$ EST PRESQUE NORMALE, ON PEUT UTILISER LA RÈGLE EMPIRIQUE QUI DIT QU'APPROXIMATIVEMENT 68 % DE NOS ESTIMATIONS D'ÉCHANTILLON SERONT À UN ÉCART-TYPE DE LA VRAIE VALEUR p.





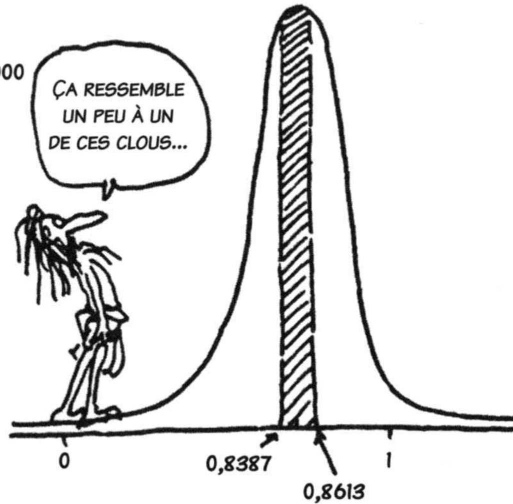
REVENONS À NOS CLOUS AVEC $n = 1000$ ET $p = 0{,}85$, L'ÉCART-TYPE EST DE :

$$\sigma(\hat{P}) = \sqrt{\frac{(0{,}85)(0{,}15)}{1000}}$$

$$= 0{,}0113$$

68 % DE NOS ESTIMATIONS PONCTUELLES DOIVENT ÊTRE DANS L'INTERVALLE

$$0{,}8387 \leq \hat{p} \leq 0{,}8613$$



L'ÉCART-TYPE DE $\hat{P}$ EST UNE MESURE DE L'**erreur-type d'échantillonnage.** COMME NOUS L'AVONS VU, CETTE ERREUR D'ÉCHANTILLONNAGE EST INVERSEMENT PROPORTIONNELLE À $\sqrt{n}$. AUGMENTER LA TAILLE DE L'ÉCHANTILLON D'UN FACTEUR 4 RÉDUIT L'ERREUR-TYPE $\sigma(\hat{P})$ DE MOITIÉ.

DÉJÀ, AVEC $n = 100$, ON VOIT QUE $\sigma(\hat{p})$ EST DE L'ORDRE DE $3\frac{1}{2}$ %.



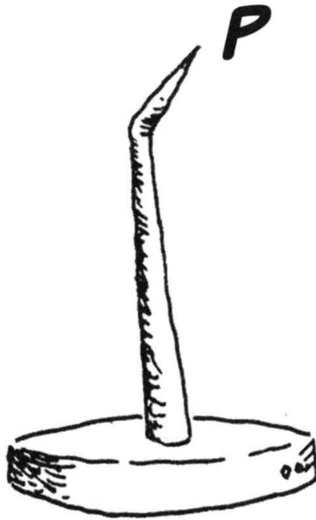
TAILLE D'ÉCHANTILLON DES CLOUS, $p = 0{,}85$

n	1	4	16	25	100	10 000
$\sqrt{n}$	1	2	4	5	10	100
$\sigma(\hat{p})$	0,357	0,1785	0,089	0,071	0,0357	0,0036

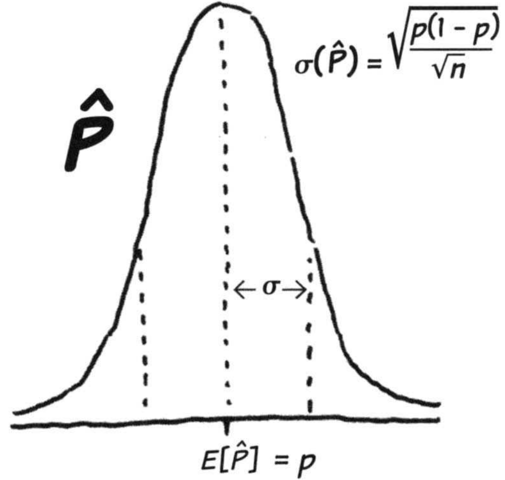
NOTE DE VOCABULAIRE : UNE **ESTIMATION** OU ESTIMATION PONCTUELLE EST UNE MESURE OBSERVÉE SUR UN ÉCHANTILLON. UN **ESTIMATEUR** EST UNE RÈGLE POUR OBTENIR CES ESTIMATIONS. L'ESTIMATEUR EST LA VARIABLE ALÉATOIRE $\hat{P} = X/n$.

LES STATISTIQUES IMPLIQUENT EN GÉNÉRAL UN PROCESSUS EN 4 ÉTAPES COMME NOUS VENONS DE LE FAIRE :

DÉFINIR LA POPULATION AVEC LE PARAMÈTRE INCONNU.



TROUVER UN ESTIMATEUR, SA DISTRIBUTION D'ÉCHANTILLONNAGE THÉORIQUE ET SON ERREUR-TYPE.



PRENDRE UN VRAI ÉCHANTILLON ALÉATOIRE ET TROUVER UNE ESTIMATION PONCTUELLE.



INDIQUER LE RÉSULTAT ET L'ERREUR STATISTIQUE D'ÉCHANTILLONNAGE.



Distribution d'échantillonnage de la MOYENNE

MAINTENANT, ON PASSE DES CLOUS EN LAITON AUX CORNICHONS...



LES FABRICANTS DE BOCAUX VEULENT CONNAÎTRE LA **TAILLE MOYENNE** DES CORNICHONS SANS AVOIR À INSPECTER TOUTES LES CUCURBITACÉES DE LA FRANCE. ILS SÉLECTIONNENT AU HASARD *n* CORNICHONS ET MESURENT LEURS TAILLES $x_1, x_2 \dots x_n$.

À PRÉSENT, VOUS DEVEZ VOUS DOUTER QUE CHAQUE X_i EST UNE **VARIABLE ALÉATOIRE** EN TANT QUE RÉSULTAT D'UNE EXPÉRIENCE ALÉATOIRE.



SI μ EST LA MOYENNE (INCONNUE) ET σ EST L'ÉCART-TYPE DE LA **DISTRIBUTION DES TAILLES DE CORNICHONS,** ALORS :

$$E[X_i] = \mu$$
$$\sigma[X_i] = \sigma$$

POUR CHAQUE i (CAR x_i EST UNE TAILLE POSSIBLE DE CORNICHONS).

C'EST ÉTRANGE TOUT CE QU'ON SAIT SUR LES VARIABLES ALÉATOIRES, ALORS QU'IL Y A SEULEMENT UNE MINUTE ON NE SAVAIT PAS CE QU'ÉTAIT UNE VARIABLE ALÉATOIRE...



MAINTENANT REGARDONS LA MOYENNE D'ÉCHANTILLON, CELLE DES CORNICHONS SÉLECTIONNÉS. C'EST UNE NOUVELLE VARIABLE ALÉATOIRE DONNÉE PAR :

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$



COMME AUPARAVANT, NOUS VOUDRIONS CONNAÎTRE À QUEL POINT CETTE VALEUR APPROCHE μ. AUTREMENT DIT, SI ON DISPOSAIT DE PLUSIEURS ÉCHANTILLONS DIFFÉRENTS, QUELLE SERAIT LA DISTRIBUTION DE $\bar{X}$? COMME NOUS CONNAISSONS X_1, X_2 ... ET X_n, ON SAIT AUSSI QUE :

$$E[\bar{X}] = \mu$$

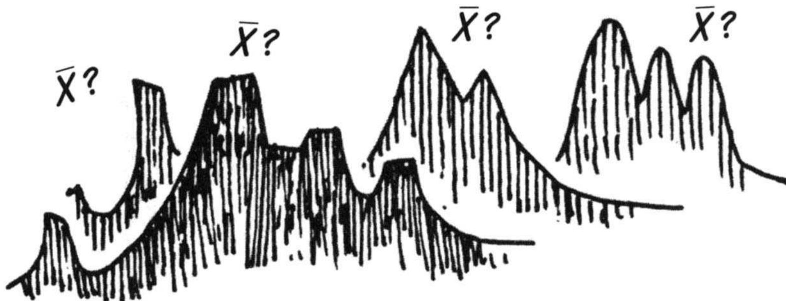
$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

À NOUVEAU, ON RETROUVE NOTRE DÉNOMINATEUR MAGIQUE ! L'ÉCART-TYPE DES OBSERVATIONS D'ÉCHANTILLONS DÉPEND DE

$$\frac{1}{\sqrt{n}}$$

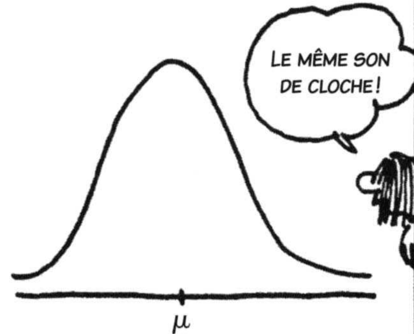


MAIS NOUS NE CONNAISSONS PAS LA FORME DE LA DISTRIBUTION DE $\bar{X}$. LA **DISTRIBUTION D'ÉCHANTILLONNAGE** DE $\hat{P}$ ÉTAIT PROCHE D'UNE NORMALE, CAR BASÉE SUR UNE VARIABLE ALÉATOIRE BINOMIALE. MAIS QU'EN EST-IL DE $\bar{X}$, L'ESTIMATEUR DE LA **MOYENNE** D'ÉCHANTILLON ?



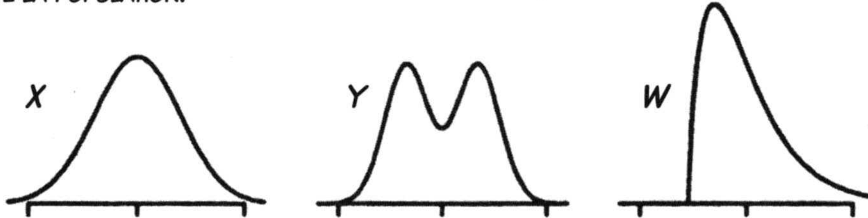
IL SE TROUVE QUE $\bar{X}$ EST ÉGALEMENT PRESQUE NORMALE ! CE RÉSULTAT BIEN CONNU PORTE AUSSI LE NOM DE **THÉORÈME CENTRAL LIMITE.**

IL DIT QUE SI L'ON PREND DES ÉCHANTILLONS ALÉATOIRES DE TAILLE n D'UNE POPULATION DE MOYENNE μ ET D'ÉCART-TYPE σ, ALORS PLUS n AUGMENTE, PLUS $\bar{X}$ TEND VERS UNE **DISTRIBUTION NORMALE** DE MOYENNE μ ET D'ÉCART-TYPE $\sigma/\sqrt{n}$. AINSI,

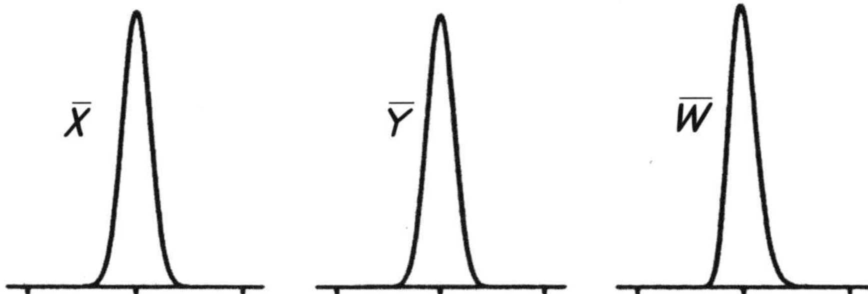


$$P(a \leq \bar{X} \leq b) \approx P\left(\frac{a - \mu}{\sigma/\sqrt{n}} \leq Z \leq \frac{b - \mu}{\sigma/\sqrt{n}}\right)$$

POURQUOI EST-CE REMARQUABLE ? CELA INDIQUE QUE, QUELLE QUE SOIT LA FORME DE LA DISTRIBUTION INITIALE (DES TAILLES DE CORNICHONS POUR NOUS), LES **MOYENNES** D'ÉCHANTILLON TENDENT VERS UNE LOI **NORMALE.** POUR CONNAÎTRE LA DISTRIBUTION DE $\bar{X}$, IL NOUS SUFFIT DE CONNAÎTRE LA MOYENNE ET L'ÉCART-TYPE DE LA POPULATION.



LES TROIS DENSITÉS PRÉCÉDENTES ONT LES MÊMES MOYENNES ET ÉCARTS-TYPES. MALGRÉ LEURS FORMES DIFFÉRENTES, LORSQUE $n = 10$, LES DISTRIBUTIONS D'ÉCHANTILLONNAGE DE LA MOYENNE $\bar{X}$ SONT QUASIMENT IDENTIQUES.



Le « *t* de Student »

AUSSI INCROYABLE QUE SOIT LE THÉORÈME CENTRAL LIMITE, IL A AU MOINS DEUX DÉFAUTS.



LE PREMIER : IL FAUT QUE n SOIT GRAND.

LE SECOND : IL FAUT CONNAÎTRE L'ÉCART-TYPE σ.

MAIS LES ÉCHANTILLONS SONT SOUVENT PETITS ET σ EST GÉNÉRALEMENT INCONNU. QUOI QU'IL EN SOIT DANS LE CAS DES CORNICHONS, NOUS N'AVONS **AUCUNE IDÉE** DE LA DISPERSION DES TAILLES DES CORNICHONS AUTOUR DE LA MOYENNE.



ON PEUT ALORS **ESTIMER** σ EN PRENANT **s L'ÉCART-TYPE À L'INTÉRIEUR DE L'ÉCHANTILLON** OBTENU À PARTIR DE :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

AINSI À L'INTÉRIEUR DE LA VARIABLE ALÉATOIRE :

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

ON REMPLACE σ PAR s ET ON DÉFINIT UNE NOUVELLE **VARIABLE ALÉATOIRE t** :

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$



On peut voir la variable aléatoire t comme la **meilleure approximation possible étant donné les circonstances.** Sa distribution est appelée **t de Student,** car elle a été publiée par son inventeur **William Gosset** sous le pseudonyme « Étudiant ».

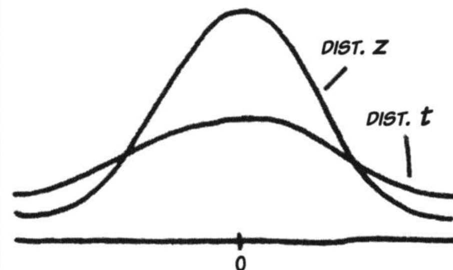


(Gosset était un employé de la **brasserie Guinness,** qui lui demanda d'utiliser un pseudonyme pour diverses raisons).

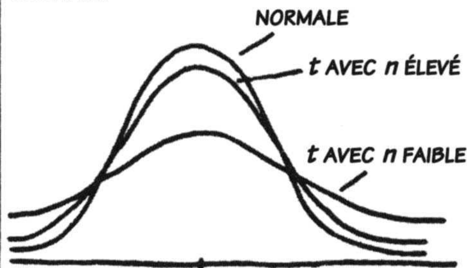
En faisant l'hypothèse que la **population originale des observations** était **normale,** ou presque, « Student » put fournir une conclusion :



Le t de Student est plus dispersé et plus aplati qu'une loi normale (z). Cela s'explique par l'utilisation de l'écart-type d'échantillon (s) qui introduit encore plus d'incertitude dans la distribution.



La quantité de dispersion dépend de n, la **taille de l'échantillon.** Plus la taille d'échantillon est élevée, plus s approxime bien σ et t se rapproche de z, la loi normale.



Gosset put établir des tables de t pour différentes tailles d'échantillon et nous les utiliserons dans le prochain chapitre.



DANS CE CHAPITRE, NOUS AVONS VU LE PROBLÈME CENTRAL DES **STATISTIQUES APPLIQUÉES :** COMMENT CONSTRUIRE UN **ÉCHANTILLON** D'UNE GRANDE POPULATION AFIN QUE LES ANALYSES STATISTIQUES SOIENT VALIDES. OUTRE LA « NORME STANDARD » DE L'ÉCHANTILLON ALÉATOIRE SIMPLE, NOUS AVONS VU D'AUTRES MÉTHODES D'ÉCHANTILLONNAGE QUI PEUVENT ÊTRE UTILISÉES POUR DES RAISONS PRATIQUES D'EFFICACITÉ OU DE COÛT.



ENSUITE, ET AVEC UN ÉCHANTILLON ALÉATOIRE SIMPLE, NOUS AVONS VU COMMENT LES STATISTIQUES D'ÉCHANTILLONS ÉTAIENT **DISTRIBUÉES.** AINSI ON A VU QUE PRENDRE UN ÉCHANTILLON CONSTITUAIT UNE **EXPÉRIENCE ALÉATOIRE,** ET QUE LES STATISTIQUES OBSERVÉES DEVENAIENT DES **VARIABLES ALÉATOIRES.**



ON A VU QUE LA VARIABLE ALÉATOIRE **PROPORTION** D'ÉCHANTILLON $\hat{p}$, ET QUE LA VARIABLE ALÉATOIRE **MOYENNE** D'ÉCHANTILLON $\overline{X}$ ÉTAIENT À PEU PRÈS DISTRIBUÉES NORMALEMENT LORSQUE LA TAILLE D'ÉCHANTILLON ÉTAIT SUFFISANTE. ON A AUSSI INTRODUIT LES VARIABLES **STANDARDS** z ET t, QUI SERVIRONT SUR LES RÉSULTATS D'ÉCHANTILLONS POUR LES CHAPITRES SUIVANTS.



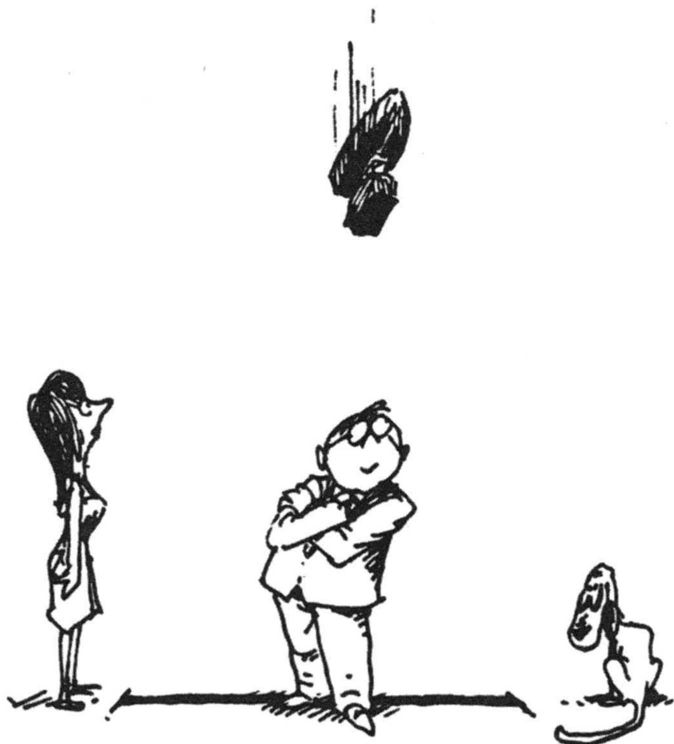
*AVEC DU THÉ BIEN SÛR !

DANS LES DEUX PROCHAINS CHAPITRES, NOUS ALLONS VOIR COMMENT UTILISER CES DISTRIBUTIONS POUR FAIRE DES **INFÉRENCES STATISTIQUES :** À PARTIR D'UN SEUL ÉCHANTILLON, UN SONDAGE D'OPINION PAR EXEMPLE, COMMENT MOBILISER NOS CONNAISSANCES SUR $\hat{P}$ ET $\bar{X}$ POUR LES ÉVALUER ?

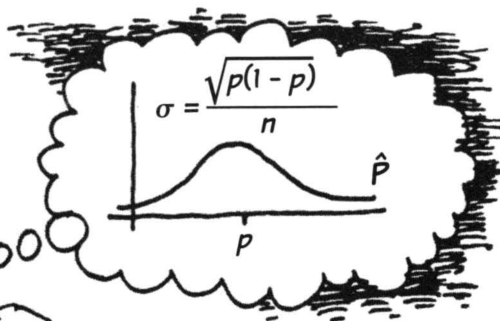


Chapitre 7

Intervalles de confiance



DANS LE CHAPITRE PRÉCÉDENT, NOUS AVONS VU LES MÉTHODES D'ÉCHANTILLONNAGE. À PARTIR D'UNE GRANDE POPULATION, NOUS AVONS ANALYSÉ COMMENT LA DISTRIBUTION D'ESTIMATEURS D'ÉCHANTILLON VARIAIT SELON LES ÉCHANTILLONS.





DANS CE CHAPITRE, NOUS FERONS L'INVERSE. À PARTIR D'UN ÉCHANTILLON, NOUS NOUS DEMANDERONS QUEL EST LE SYSTÈME ALÉATOIRE QUI A GÉNÉRÉ NOS STATISTIQUES ?



CELA CHANGE NOTRE MODE DE PENSÉE : ON PASSE D'UN RAISONNEMENT DÉDUCTIF À UN RAISONNEMENT **INDUCTIF.**





DANS UN RAISONNEMENT **DÉDUCTIF,** ON PART D'UNE HYPOTHÈSE POUR ARRIVER À UNE CONCLUSION : « SI LORD FINE-GACHETTE A COMMIS LE CRIME, ALORS IL A EFFACÉ SES EMPREINTES DE L'ARME. »

LE RAISONNEMENT **INDUCTIF** EST **À REBOURS,** ON PART D'UN ENSEMBLE D'OBSERVATIONS VERS UNE HYPOTHÈSE VRAISEMBLABLE.







À BIEN DES ÉGARDS, LA SCIENCE ET DONC LES STATISTIQUES RESSEMBLENT À UN TRAVAIL DE DÉTECTIVE. À PARTIR D'UN ENSEMBLE D'OBSERVATIONS, NOUS NOUS INTERROGEONS SUR CE QUI LES A GÉNÉRÉES.

ESTIMATION
d'intervalles de confiance

C'EST UNE FORME D'INFÉRENCE STATISTIQUE, QUE L'ON VOIT APPARAÎTRE CHAQUE FOIS PENDANT LES **PÉRIODES ÉLECTORALES...**



QUELQUE PART, LORS D'UNE ÉLECTION RÉCENTE, LE SÉNATEUR EN PLACE **ASTUTE** COMMANDE UN SONDAGE À **HOLMES RECHERCHE INSTITUT.** LE SONDEUR HOLMES PREND ALORS UN ÉCHANTILLON ALÉATOIRE SIMPLE DE 1000 ÉLECTEURS ET LEUR DEMANDE CE QU'ILS PENSENT D'ASTUTE.



APRÈS AVOIR CENSURÉ LES REMARQUES DES QUELQUES VALEURS EXTRÊMEMENT GRINCHEUSES, HOLMES TROUVE QUE **550** VOTANTS SONT EN FAVEUR DE SON CLIENT, LE SÉNATEUR ASTUTE.



CELA CONSTITUE L'ESTIMATION PONCTUELLE.


VOUS AVEZ SEULEMENT INTERROGÉ MILLE PERSONNES ? ! MAIS IL Y A PLUS D'UN MILLION DE VOTANTS DANS L'ÉTAT !
PEU IMPORTE QU'IL Y AIT UN MILLION OU UN MILLIARD DE VOTANTS...
J'AI PRIS UN ÉCHANTILLON ALÉATOIRE SIMPLE ! CELA IMPLIQUE QUE JE PEUX VOUS PROPOSER UNE GARANTIE EN BÉTON !

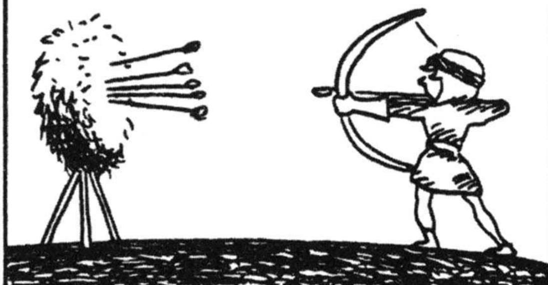

CETTE GARANTIE... ELLE EST ABSOLUE, DU TYPE 100 % REMBOURSÉE, C'EST BIEN CELA ?
HEIN ?
HUM... AH ! BIEN. C'EST STATISTIQUE EN FAIT...
OUI, JE PEUX DIRE AVEC 95 % DE CONFIANCE QUE p, LA VRAIE PROPORTION DE POPULATION, EST ENTRE 0,519 ET 0,581 !


SÛR À 95 % ? ! SUPERBE ! QUE CROYEZ-VOUS QU'IL SE PASSERAIT SI JE FAISAIS CAMPAGNE AVEC 95 % D'HONNÊTETÉ ?
JE NE SAIS PAS, PERSONNE N'A JAMAIS ESSAYÉ...
UNE FOIS ASTUTE CALMÉ, HOLMES LUI EXPLIQUE LE SENS D'UNE CONFIANCE À 95 % : IL SAIT QUE SA PROCÉDURE D'ESTIMATION A UNE PROBABILITÉ DE 95 % DE PRODUIRE UN INTERVALLE CONTENANT p. AINSI PENDANT SES NOMBREUSES ANNÉES DE SONDAGE, p ÉTAIT DANS L'INTERVALLE DE CONFIANCE AUTOUR DE p̂, LA VALEUR OBSERVÉE, DANS 95 % DES CAS.

LE SÉNATEUR ASTUTE DEMEURE PERPLEXE ! ALORS HOLMES LUI DONNE UNE **leçon de tir à l'arc.**



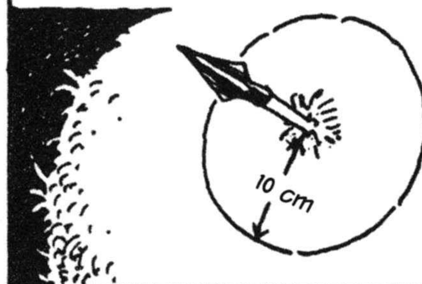
CONSIDÉRONS UN ARCHER SONDEUR QUI VISE UNE CIBLE. SUPPOSONS QU'IL TOUCHE 95 % DU TEMPS LE CENTRE DE LA CIBLE DANS UN RAYON DE 10 cm. AINSI, SEULE UNE FLÈCHE SUR 20 N'ATTEINT PAS CE DISQUE.



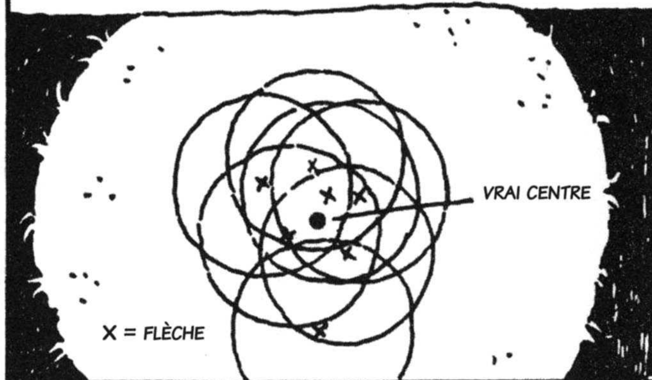
NOTRE BRAVE DÉTECTIVE EST ASSIS DERRIÈRE LA CIBLE ET N'EN VOIT PAS LE CENTRE. L'ARCHER TIRE UNE SEULE FLÈCHE.



CONNAISSANT L'AGILITÉ DU TIREUR, LE DÉTECTIVE TRACE UN CERCLE DE RAYON 10 cm AUTOUR DE LA POINTE DE LA FLÈCHE. IL EST MAINTENANT **CONFIANT À 95 %** QUE LE DISQUE INCLUT LE CENTRE DE LA CIBLE.



SON RAISONNEMENT EST QUE, S'IL TRACE DES CERCLES DE RAYON DE 10 cm AUTOUR DE **PLUSIEURS FLÈCHES,** LES DISQUES VONT INCLURE LE CENTRE DANS 95 % DES CAS.



(LES PROBABILISTES UTILISENT LE TERME **STOCHASTIQUE** POUR DÉCRIRE CES MODÈLES ALÉATOIRES.
CELA VIENT DU GREC ***STOKHAZESTHAI*** QUI VEUT DIRE VISER UNE CIBLE, ET DE ***STOKHOS*** UNE FLÈCHE.)

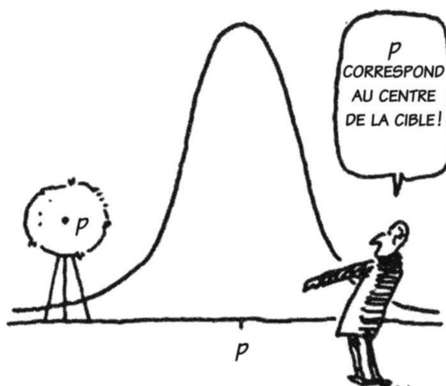




HOLMES TRADUIT MAINTENANT LA LEÇON DE TIR À L'ARC DANS LE LANGAGE DÉVELOPPÉ AU CHAPITRE PRÉCÉDENT.

Premier temps. TIREZ BEAUCOUP DE FLÈCHES.

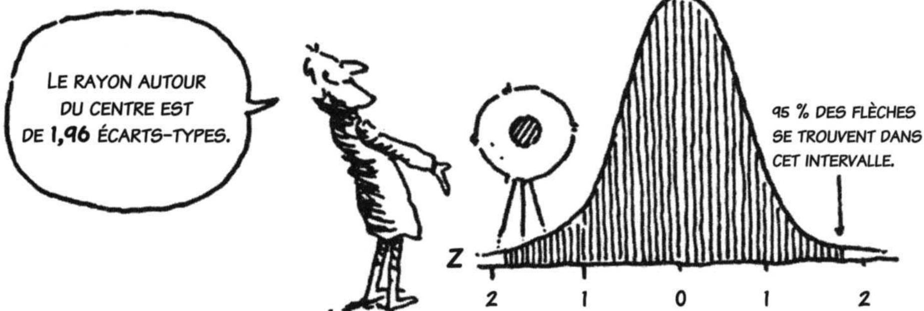
UN CALCUL DE PROBABILITÉ PERMET D'ESTIMER LA POSITION DU CENTRE DE LA CIBLE. LES ESTIMATIONS PONCTUELLES $\hat{p}$ CORRESPONDENT À NOS FLÈCHES. LA DISTRIBUTION D'ÉCHANTILLONNAGE DE $\hat{P}$ EST PRESQUE NORMALE DE MOYENNE μ ET D'ÉCART-TYPE



$$\sigma(\hat{P}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

COMME LA COURBE EST NORMALE, NOUS UTILISONS SON Z-SCORE ET UNE TABLE STANDARD POUR ÉVALUER LA LARGEUR DE L'INTERVALLE DANS LEQUEL 95 % DES FLÈCHES TOMBENT (NOUS ALLONS VOIR EXACTEMENT COMMENT CALCULER CELA DANS QUELQUES PAGES). NOUS TROUVONS QUE CETTE LARGEUR EST DE **1,96** ÉCARTS-TYPES.

$$0{,}95 = P(-1{,}96 \leq z \leq 1{,}96)$$



MAINTENANT, FAISONS UN PEU D'ALGÈBRE.
ON TRANSFORME EN LOI NORMALE STANDARD Z :



$$0{,}95 \approx P\left(-1{,}96 \leq \frac{\hat{P} - p}{\sigma(\hat{P})} \leq 1{,}96\right)$$

CE QUI SE TRADUIT PAR :

$$0{,}95 \approx P(p - 1{,}96\sigma(\hat{P}) \leq \hat{P} \leq p + 1{,}96\sigma(\hat{P}))$$

CE QUI REVIENT À DIRE QUE 95 % DES « FLÈCHES » $\hat{p}$ TOMBENT ENTRE $p - 1{,}96\sigma(\hat{P})$ ET $p + 1{,}96\sigma(\hat{P})$.

MAINTENANT, NOUS POUVONS REGARDER LA CIBLE DE DERRIÈRE. ENCORE UN TOUR DE MANIVELLE ALGÉBRIQUE ET ON OBTIENT :

$$0{,}95 \approx P(\hat{P} - 1{,}96\sigma(\hat{P}) \leq p \leq \hat{P} + 1{,}96\sigma(\hat{P}))$$

ICI, ON DESSINE LES CERCLES AUTOUR DES POINTES DE FLÈCHE, AINSI ON TRACE DES INTERVALLES AUTOUR DE $\hat{p}$, ET ON DIT QUE 95 % D'ENTRE EUX COUVRENT p.



MAIS IL Y A UN LÉGER PROBLÈME : ON NE CONNAÎT PAS EXACTEMENT LA **TAILLE** AUTOUR DE LA CIBLE, CAR SA LARGEUR EST UN MULTIPLE DE $\sigma(\hat{P})$ QUI DÉPEND DE L'INCONNU p.



ON ESQUIVE DONC LE PROBLÈME EN UTILISANT L'ESTIMATION PONCTUELLE $\hat{p}$ À LA PLACE DE p POUR DÉTERMINER L'**ERREUR-TYPE** ESTIMÉE OU ERREUR-STANDARD :

$$s(\hat{P}) = \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}}$$

C'EST SUFFISAMMENT PROCHE... C'EST AU MIEUX... ET CELA PEUT SE JUSTIFIER THÉORIQUEMENT !

MAINTENANT LA FORMULE DEVIENT :

$$0{,}95 \approx P(\hat{P} - 1{,}96s(\hat{P}) \leq p \leq \hat{P} + 1{,}96s(\hat{P}))$$

DE NOUVEAU, CETTE ÉQUATION DÉCRIT LA PROBABILITÉ QUE LA VRAIE VALEUR DE PROPORTION DE LA POPULATION SOIT DANS L'**INTERVALLE ALÉATOIRE.**

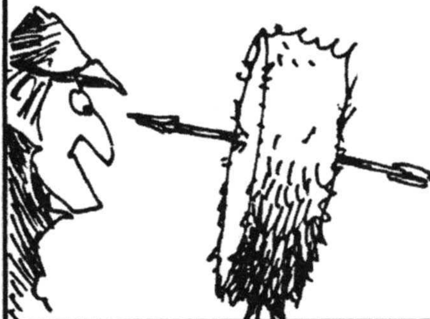


$$[\hat{P} - 1{,}96s(\hat{P}), \hat{P} + 1{,}96s(\hat{P})]$$

SI L'ON PREND DES ÉCHANTILLONS RÉPÉTÉS, LES INTERVALLES $[\hat{p} - 1{,}96s(\hat{P}), \hat{p} + 1{,}96s(\hat{P})]$ COUVRIRONT p DANS 95 % DES CAS.

MAINTENANT QUE LES CALCULS SONT FAITS, IL EST TEMPS DE PASSER AU...

Second temps. LE TRAVAIL DE DÉTECTIVE. DANS UN VRAI SONDAGE, HOLMES PREND UN SEUL ÉCHANTILLON DE 1000 VOTANTS, TROUVE $\hat{p} = 0{,}55$ ET VEUT EN DÉDUIRE p.



IL UTILISE NOTRE PREMIER TEMPS POUR CALCULER :

$$s(\hat{P}) = \frac{\sqrt{(0{,}55) \times (0{,}45)}}{\sqrt{1000}} = 0{,}0157$$

ET CONCLUT QU'AVEC 95 % DE CONFIANCE p EST DANS L'INTERVALLE :

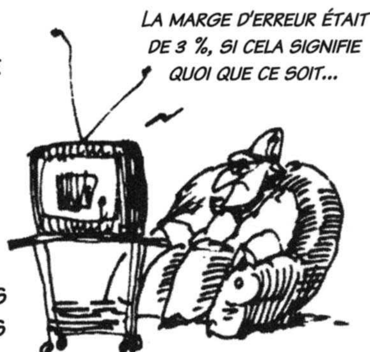
$$\hat{p} \pm 1{,}96s(\hat{P})$$

$$= 0{,}550 \pm 1{,}96 \times 0{,}0157$$

$$= 0{,}550 \pm 0{,}031$$

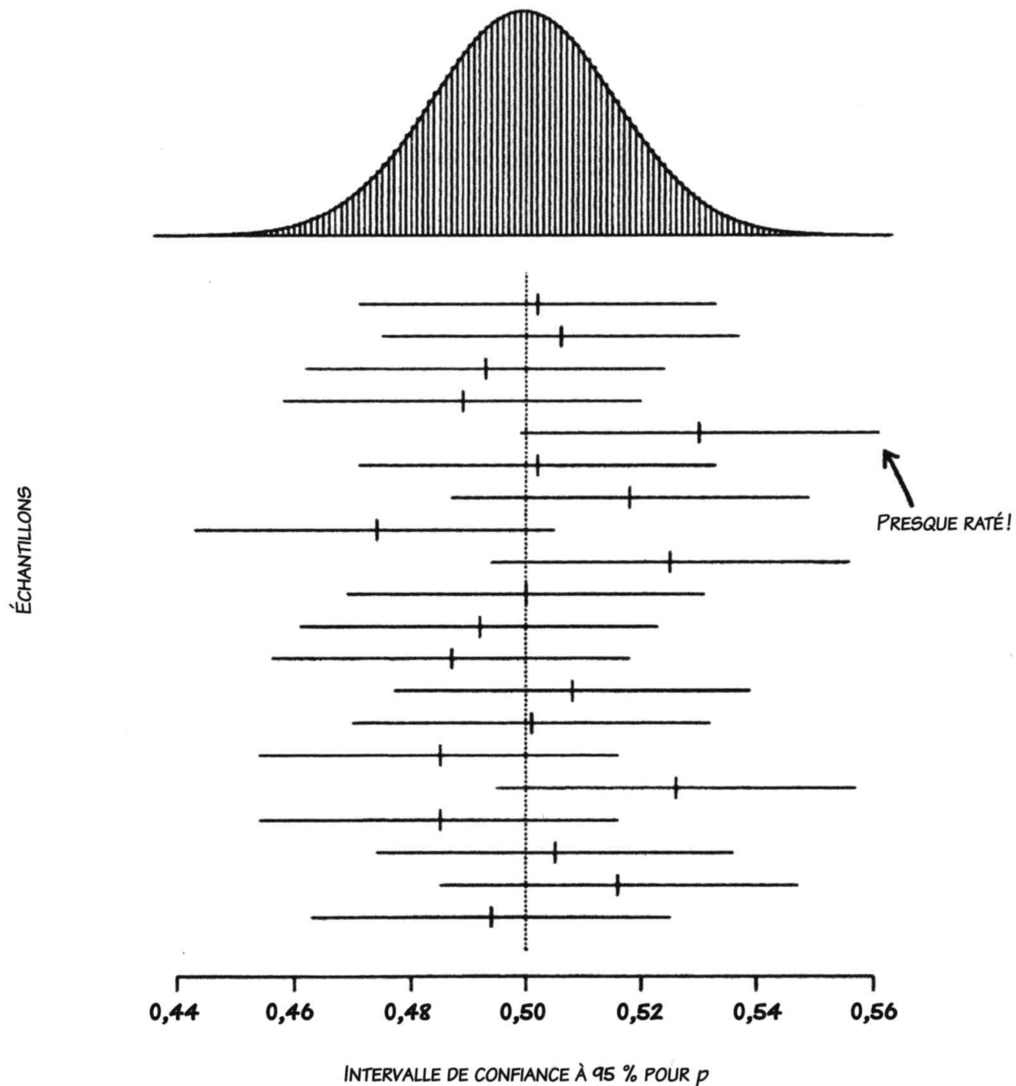
C'EST CE À QUOI FONT RÉFÉRENCE LES SONDAGES EN PARLANT DE « MARGE D'ERREUR ». DANS NOTRE CAS, HOLMES A TROUVÉ QUE :

$$0{,}519 \leq p \leq 0{,}581$$



EN D'AUTRES TERMES, p = 55 % AVEC 3 % DE MARGE D'ERREUR (LES SONDAGES UTILISENT TYPIQUEMENT DES INTERVALLES À 95 %).

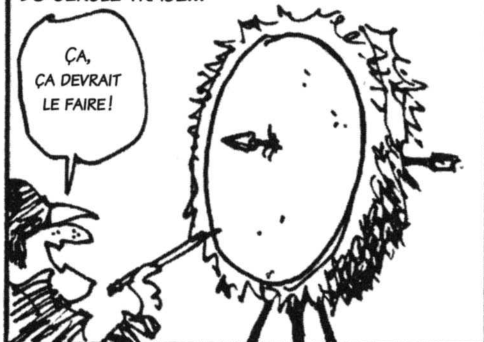
NOUS MONTRONS SUR CETTE PAGE LES RÉSULTATS D'UNE SIMULATION PAR ORDINATEUR DE VINGT ÉCHANTILLONS DE TAILLE $n = 1000$. LA SIMULATION UTILISE LA VALEUR $p = 0{,}5$. LA DISTRIBUTION D'ÉCHANTILLONNAGE DE $\hat{P}$ (NORMALE DE MOYENNE p ET ÉCART-TYPE $\sigma = \sqrt{p(1-p)/n}$) APPARAÎT EN HAUT. EN DESSOUS, IL Y A LES INTERVALLES DE CONFIANCE À 95 % DE CHAQUE ÉCHANTILLON. UN SUR VINGT (SOIT 5 %) DE CES INTERVALLES **NE COUVRIRONT PAS LA VALEUR** p = 0,5.



BIEN QUE 95 % DE CONFIANCE SUFFISENT POUR LES SONDAGES DES JOURNAUX, CE N'EST PAS ASSEZ POUR LE SÉNATEUR ASTUTE. IL VEUT **99 %** !



COMMENT AUGMENTER LA CONFIANCE ? PAR ANALOGIE AVEC LE TIR À L'ARC, DEUX MÉTHODES SONT POSSIBLES : LA PREMIÈRE CONSISTE À **AUGMENTER LA TAILLE** DU CERCLE TRACÉ...



ET UNE AUTRE FAÇON SERAIT D'**AMÉLIORER LA VISÉE** INITIALE DE L'ARCHER, DE MANIÈRE QUE LE TIR DES FLÈCHES SOIT PLUS GROUPÉ AU CENTRE DE LA MIRE.



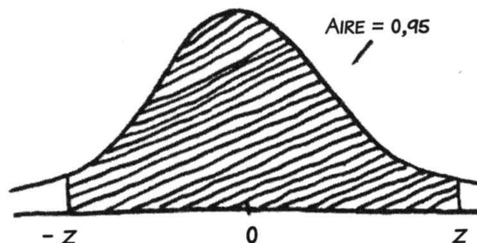
LA PREMIÈRE MÉTHODE REVIENT À **ÉLARGIR L'INTERVALLE DE CONFIANCE.** PLUS LA MARGE D'ERREUR EST IMPORTANTE, PLUS ON EST SÛR QUE LA VRAIE VALEUR p SERA DANS L'INTERVALLE.



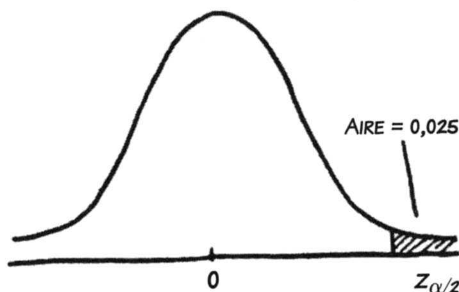
PEUT-ÊTRE EST-IL TEMPS DE VOIR EXACTEMENT COMMENT ON DÉTERMINE LES BORNES DE CES INTERVALLES DE CONFIANCE...

ICI, LE NOMBRE PERTINENT S'APPELLE α (ALPHA). IL MESURE LA DIFFÉRENCE ENTRE LE SEUIL DE CONFIANCE VOULU ET LA CERTITUDE. PAR EXEMPLE, QUAND LE SEUIL DE CONFIANCE VAUT 95 %, OU 0,95, ALORS α = **0,05**. ON PARLE DONC D'UN SEUIL (OU NIVEAU) DE CONFIANCE DE $(1 - \alpha)$.

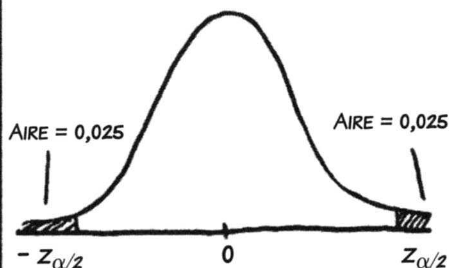
CHERCHER UN INTERVALLE DE CONFIANCE À $(1 - \alpha)$ REVIENT À REGARDER LA LOI NORMALE ET À CHERCHER LES POINTS $\pm z$ ENTRE LESQUELS L'AIRE VAUT $(1 - \alpha)$.



LE POINT $z_{\alpha/2}$, APPELÉ **FRACTILE**, EST LA VALEUR Z AU-DESSUS DE LAQUELLE L'AIRE EST DE $\alpha/2 = 0{,}025$.



ON ENLÈVE LES QUEUES DE DISTRIBUTION AUX EXTRÉMITÉS DE LA COURBE, QUI ONT UNE AIRE TOTALE DE $\alpha = \alpha/2 + \alpha/2$.



Z	- 2,5	- 2,4	- 2,3	- 2,2	- 2,1	- 2,0	- 1,9	- 1,8	- 1,7	- 1,6	- 1,5	- 1,4
F(Z)	0,006	0,008	0,011	0,014	0,018	0,023	0,029	0,036	0,045	0,055	0,067	0,081

ON PEUT TROUVER $z_{\alpha/2}$ DIRECTEMENT AVEC LA TABLE DE LOI NORMALE DE LA PAGE 84. IL S'AGIT DU POINT AYANT CETTE PROPRIÉTÉ :

$$P(z \geq z_{\alpha/2}) = \frac{\alpha}{2}$$

EN PARTICULIER :

$$P(z \geq z_{0{,}025}) = 0{,}025$$



VOICI UNE PETITE TABLE DE FRACTILES POUR DIFFÉRENTS SEUILS DE CONFIANCE :

$1-\alpha$	0,8	0,9	0,95	0,99
α	0,2	0,1	0,05	0,01
$\alpha/2$	0,1	0,05	0,025	0,005
$z_{\alpha/2}$	**1,28**	**1,64**	**1,96**	**2,58**



POUR FAIRE UN INTERVALLE DE CONFIANCE À 99 %, ON UTILISE CETTE TABLE POUR ÉCRIRE :

$$0{,}99 = P(\hat{p} - 2{,}58\,s(\hat{P}) \leq p \leq \hat{p} + 2{,}58\,s(\hat{P}))$$

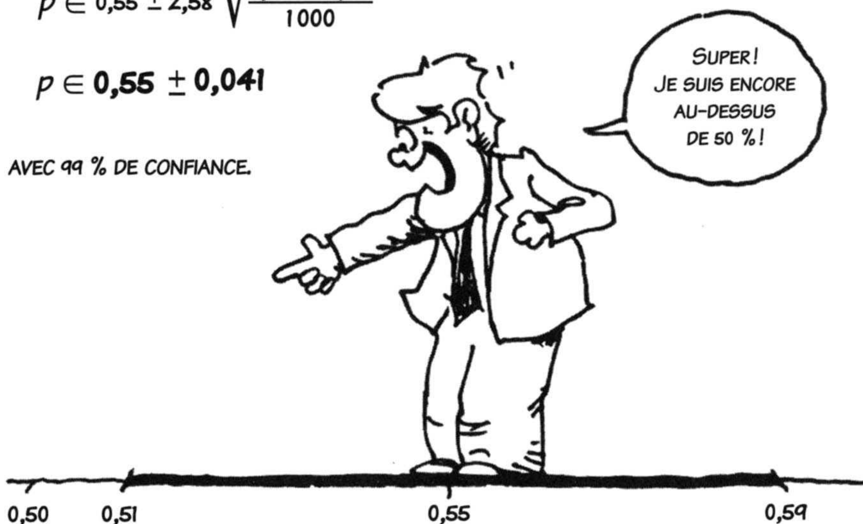
QUE NOUS ABRÉGEONS :

$$p \in \hat{p} \pm 2{,}58\,s(\hat{P})$$

$$p \in 0{,}55 \pm 2{,}58\sqrt{\frac{0{,}55 \times 0{,}45}{1000}}$$

$$p \in \mathbf{0{,}55 \pm 0{,}041}$$

AVEC 99 % DE CONFIANCE.

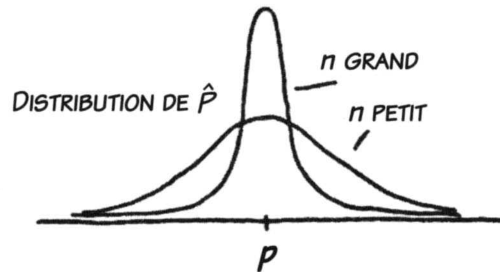


AUGMENTER LES INTERVALLES EST UNE FAÇON D'AUGMENTER LA CONFIANCE DANS SES RÉSULTATS. COMME NOUS L'AVONS DIT, UNE AUTRE FAÇON SERAIT DE VISER PLUS JUSTE AVEC NOS FLÈCHES. SI NOUS SAVIONS QUE 95 % DE NOS FLÈCHES SONT À 1 cm DE LA CIBLE, NOS ESTIMATIONS POURRAIENT ÊTRE PLUS AFFÛTÉES.



COMMENT POUVONS-NOUS FAIRE ? EN AUGMENTANT LA TAILLE DE L'ÉCHANTILLON ! LA LARGEUR DE L'INTERVALLE DE CONFIANCE DÉPEND DE CETTE TAILLE : L'INTERVALLE EST DE LA FORME $\hat{p} \pm E$ OÙ L'ERREUR E EST DONNÉE PAR :

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$



PLUS n EST GRAND, PLUS L'ERREUR EST PETITE (PAR EXEMPLE QUADRUPLER n RÉDUIT DE MOITIÉ LA LARGEUR DE L'INTERVALLE).

ASTUTE DEMANDE À HOLMES UNE PETITE ERREUR AVEC UNE GRANDE CONFIANCE – DISONS 99 % DE CONFIANCE AVEC $E = \pm 0{,}01$. HOLMES CALCULE DONC n :

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2}$$

ICI p^* EST UNE PRÉVISION DE LA VRAIE PROPORTION p (RAPPELEZ-VOUS QUE CELA SE FAIT AVANT LA MESURE D'ÉCHANTILLON).



En prenant une prévision prudente de $p^* = 0{,}5$, Holmes trouve :

$$n = \frac{(2{,}58)^2(0{,}5)^2}{(0{,}01)^2}$$

$$= \frac{(6{,}65)(0{,}25)}{(0{,}0001)}$$

$$= \mathbf{16\,641}$$

1 000 votants donnaient une erreur de 3 % avec 95 % de confiance. Pour avoir 1 % d'erreur avec 99 % de confiance, Holmes doit sonder 16 641 votants.







Ils font alors le sondage et vont à l'élection avec **99 % de confiance.**

Mais... toutes ces probabilités sont utiles avant les élections. Après l'élection, le sénateur est soit 100 % élu, soit 100 % perdant... et malgré tout, le sénateur Astute **perd** l'élection.





IL S'EST PASSÉ QUE LES POLITICIENS NE SONT PAS ÉLUS AVEC LES SONDAGES !



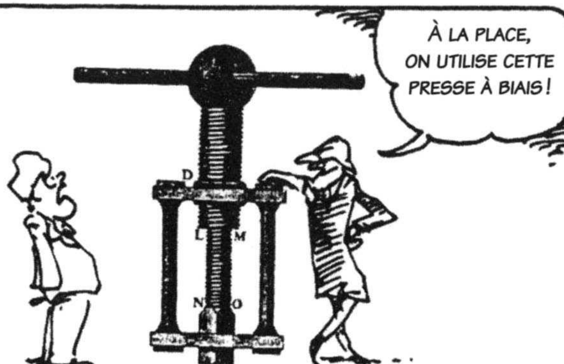
QUELQUES PROBLÈMES DES SONDAGES PAR RAPPORT AUX ÉLECTIONS :



IL N'Y A **AUCUN MOYEN** POUR UN SONDEUR D'ENTRER DANS LA TÊTE D'UN VOTANT, DE SAVOIR S'IL VA VOTER, S'IL MENT OU S'IL VA CHANGER D'AVIS AVANT L'ÉLECTION. DE GRANDS ÉCHANTILLONS NE PEUVENT PAS RÉDUIRE CE TYPE D'ERREUR.

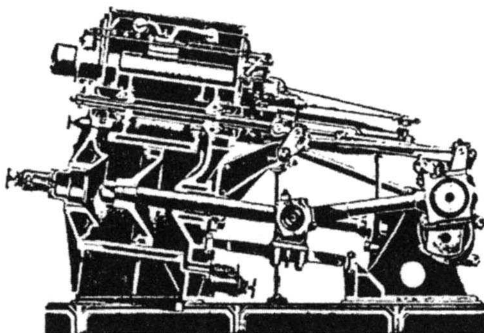


Comme ces erreurs peuvent être importantes, on gagne rarement à prendre de très grands échantillons.



Dans cinq élections présidentielles consécutives, les sondages Gallup ont interviewé moins de **4 000** votants pour chaque élection. Pourtant dans ces cinq élections, l'erreur de prédiction du résultat calculé par Gallup était de moins de **2 %**.





Ce succès est dû à l'utilisation d'estimateurs qui prennent en compte les non-réponses et filtrent les électeurs éligibles qui n'iront vraisemblablement pas voter.



En résumé, proportion estimée = vraie proportion + biais + erreur aléatoire d'échantillon. Même les sondeurs ont des ressources limitées. Ils choisissent judicieusement de dépenser de l'argent pour réduire les biais plutôt que d'augmenter le nombre de votants sondés à plus de 4 000.

Intervalle de confiance pour μ

JUSQU'À PRÉSENT, NOUS AVONS EXAMINÉ LES INTERVALLES DE CONFIANCE POUR UNE PROPORTION p D'UNE POPULATION. LE MÊME TYPE DE RAISONNEMENT FONCTIONNE POUR UNE MOYENNE μ DE POPULATION.



DANS LE CHAPITRE PRÉCÉDENT, NOUS AVONS VU PAGE 105 QUE LA DISTRIBUTION D'ÉCHANTILLONNAGE DES MOYENNES $\bar{X}$ EST PRESQUE NORMALE, CENTRÉE SUR LA MOYENNE DE POPULATION μ ET D'ÉCART-TYPE $\sigma/\sqrt{n}$ OÙ σ EST L'ÉCART-TYPE DE LA POPULATION. AINSI, SI n EST GRAND :

$$0{,}95 = P(-1{,}96 \leq z \leq 1{,}96)$$

$$\approx P\left(-1{,}96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1{,}96\right)$$



À NOUVEAU, NE CONNAISSANT PAS σ, ON REMPLACE σ PAR s, L'**ÉCART-TYPE ÉCHANTILLON**, ET ON OBTIENT :

$$0{,}95 \approx P\left(-1{,}96 \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq 1{,}96\right)$$

LE TERME $s/\sqrt{n}$ EST APPELÉ L'**ERREUR-TYPE ESTIMÉE** ET SE NOTE $s(\bar{X})$. ON EN CONCLUT QUE :

$$0{,}95 \approx P(\bar{X} - 1{,}96s(\bar{X}) \leq \mu \leq \bar{X} + 1{,}96s(\bar{X}))$$

OÙ :

$$s(\bar{X}) = \frac{s}{\sqrt{n}}$$



COMME PRÉCÉDEMMENT, NOUS AVONS TROUVÉ QUE L'INTERVALLE ALÉATOIRE

$$\overline{X} \pm 1{,}96s(\overline{X})$$

RECOUVRE LA VRAIE MOYENNE μ AVEC UNE PROBABILITÉ DE 0,95... DONC MAINTENANT ON PEUT FAIRE APPEL À SHERLOCK HOLMES POUR EFFECTUER L'**INFÉRENCE STATISTIQUE** FONDÉE SUR UN **SEUL** ÉCHANTILLON DE TAILLE n ET DE MOYENNE $\overline{x}$.

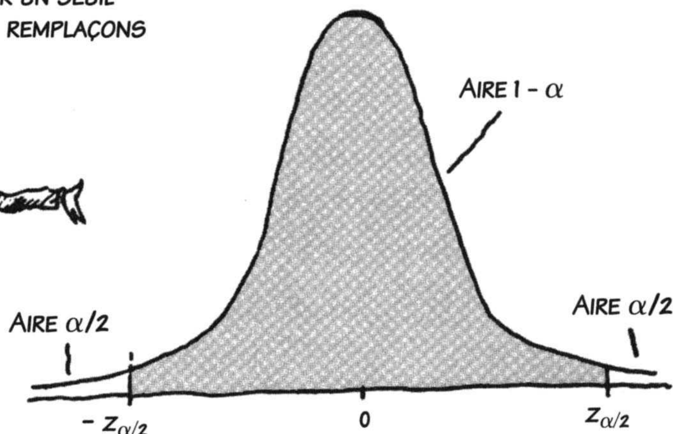


LUI ET NOUS SOMMES SÛRS À 95 % QUE LA MOYENNE μ EST DANS L'INTERVALLE $\overline{x} \pm 1{,}96s(\overline{X})$.



COMME AUPARAVANT, POUR UN SEUIL DE CONFIANCE $1 - \alpha$, NOUS REMPLAÇONS 1,96 PAR $z_{\alpha/2}$.





Revisitons les données des étudiants du chapitre 2. Supposons que nos $n = 92$ étudiants représentent un échantillon aléatoire simple de l'ensemble des étudiants de Penn State.



La moyenne d'échantillon $\bar{X}$ était de 145,20 livres et l'écart-type échantillon s était de 23,7.
Cela fait une erreur-type estimée de :

$$s(\bar{X}) = \frac{23,7}{\sqrt{92}} = 2,47$$

Et nous sommes confiants à 95 % que le poids moyen de **TOUS** les étudiants de Penn State est dans l'intervalle :

$$\bar{x} \pm 1,96s(\bar{X})$$

$$= 145,2 \pm (1,96)(2,47)$$

$$= 145,2 \pm 4,80 \text{ livres } (65,9 \pm 2,2 \text{ kg})$$

En résumé : pour un échantillon aléatoire simple de grande taille, l'intervalle de confiance à $1 - \alpha$ est :

Moyenne de population, μ

$$\mu \in \bar{x} \pm z_{\alpha/2} s(\bar{X})$$

où

$$s(\bar{X}) = \frac{s}{\sqrt{n}}$$

Proportion de population, p

$$p \in \hat{p} \pm z_{\alpha/2} s(\hat{P})$$

où

$$s(\hat{P}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

La largeur des deux intervalles est contrôlée par le seuil de confiance $1 - \alpha$ et la taille d'échantillon n.



t de Student (encore !)

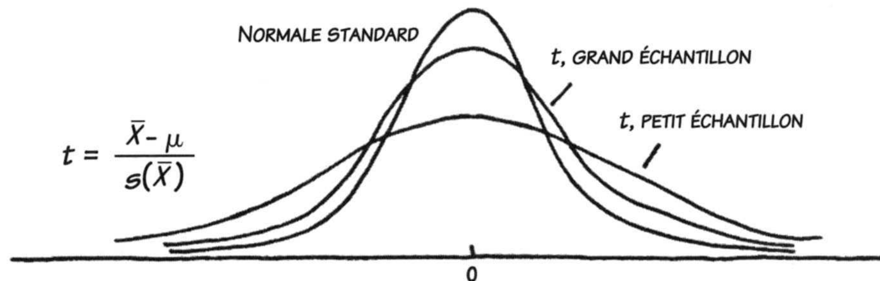
COMME NOUS L'AVONS VU DANS LE CHAPITRE 6, LA STATISTIQUE

$$\frac{\bar{X} - \mu}{s(\bar{X})}$$

EST PRESQUE NORMALE LORSQUE LA TAILLE D'ÉCHANTILLON EST GRANDE. POUR DES PETITS ÉCHANTILLONS (n = 5, 10, 25...) CE N'EST PLUS LE CAS ET NOUS DEVONS UTILISER LE t DE STUDENT.



REGARDONS LE t DE STUDENT D'UN PEU PLUS PRÈS. NOUS AVONS MENTIONNÉ QUE SA DISTRIBUTION EST PLUS ÉTALÉE QU'UNE DISTRIBUTION NORMALE, ET QUE LA DISPERSION DÉPEND DE LA TAILLE D'ÉCHANTILLON.



LA DÉCOUVERTE DE GOSSET FUT DE **QUANTIFIER** CETTE RELATION. SI n EST LA TAILLE D'ÉCHANTILLON, IL DÉFINIT $n - 1$ COMME LE NOMBRE DE **degrés de liberté** DE L'ÉCHANTILLON.

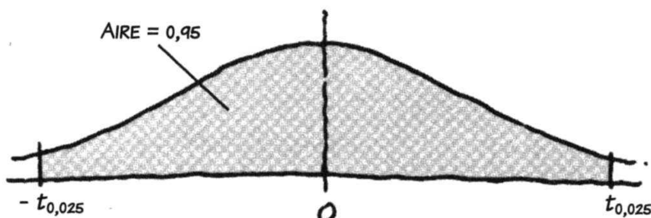


Gosset calcula des tables de la distribution t pour différentes tailles d'échantillon, et donc différents degrés de liberté. Rappelons que plus les **degrés de liberté augmentent,** plus le t de Student tend vers une loi **normale standard.**



Connaissant la taille d'échantillon n, on prend le t de Student avec $n - 1$ degrés de liberté.

Comme pour la loi normale, nous avons un seuil de confiance à 95 % en cherchant le fractile $t_{0,025}$ au-delà duquel l'aire sous la courbe est de 0,025.





Pour un intervalle de confiance à $(1 - \alpha)$, on trouve le fractile $t_{\alpha/2}$ tel que $P(t \geq t_{\alpha/2}) = \alpha/2$. Voici une table réduite de fractiles pour le t de Student.

	$1 - \alpha$	0,8	0,9	0,95	0,99
	α	0,2	0,1	0,05	0,01
	$\alpha/2$	0,1	0,05	0,025	0,005
Degrés de liberté	1	3,08	6,31	12,71	63,66
	10	1,37	1,81	2,23	3,17
	30	1,31	1,70	2,04	2,75
	100	1,29	1,66	1,98	2,63
	∞	1,28	1,64	1,96	2,58

CHAQUE COLONNE REPRÉSENTE UN SEUIL DE CONFIANCE FIXÉ EN FONCTION D'UN NOMBRE CROISSANT DE DEGRÉS DE LIBERTÉ. PLUS LE NOMBRE DE DEGRÉS DE LIBERTÉ AUGMENTE, PLUS LE FRACTILE TEND VERS $z_{\alpha/2}$, LE FRACTILE DE LA LOI NORMALE.

LA LARGEUR DE NOTRE INTERVALLE DE CONFIANCE SE DÉDUIT DE LA DÉFINITION DE t.

$$t = \frac{\bar{X} - \mu}{s(\bar{X})}$$

DONC AU SEUIL DE CONFIANCE $(1 - \alpha)$,

$$1 - \alpha = P\left(\bar{x} - t_{\alpha/2} s(\bar{X}) \leq \mu \leq \bar{x} + t_{\alpha/2} s(\bar{X})\right)$$



NOUS EN DÉDUISONS : SI L'ÉCHANTILLON EST DE TAILLE n ET DE MOYENNE $\bar{x}$, ALORS NOUS SOMMES SÛRS À $(1 - \alpha)$ QUE LA MOYENNE DE POPULATION μ APPARTIENT À L'INTERVALLE :

$$\mu \in \bar{x} \pm t_{\alpha/2} s(\bar{X})$$

OÙ $s(\bar{X}) = s/\sqrt{n}$, ET $t_{\alpha/2}$ EST LE FRACTILE DU t DE STUDENT AVEC $n - 1$ DEGRÉS DE LIBERTÉ.





NOTE :

POUR ÊTRE PLUS RIGOUREUX, L'UTILISATION DU t DE STUDENT DÉPENDAIT DE L'HYPOTHÈSE QUE LA DISTRIBUTION DE LA POPULATION ÉTAIT NORMALE. DANS LA PRATIQUE, LES INTERVALLES DE CONFIANCE BASÉS SUR LE t DE STUDENT RESTENT RELATIVEMENT VALABLES, MÊME SI LA POPULATION N'EST QU'APPROXIMATIVEMENT EN FORME DE CLOCHE.

Exemple : L'ENTREPRISE **CAMÉLÉON AUTOMOBILES** DOIT EFFECTUER DES CRASH-TESTS SUR SES VOITURES AFIN D'ÉVALUER LE COÛT MOYEN DE RÉPARATION D'UNE COLLISION FRONTALE À 16 km/h. C'EST TRÈS COÛTEUX! ELLE DÉCIDE DONC DE FAIRE UN TEST SUR SEULEMENT **CINQ** CAMÉLÉON.



ON OBTIENT LES DONNÉES DE DOMMAGES : 150 €, 400 €, 720 €, 500 € ET 930 €.

HUM! ÇA AMÉLIORE LE DESIGN.

LA MOYENNE D'ÉCHANTILLON EST :

$\bar{x}$ **= 540 €**

L'ÉCART-TYPE EST :

s = 299 €

ON PEUT VÉRIFIER S À LA CALCULATRICE. C'EST :

$$\sqrt{\frac{1}{4}\left((150-540)^2 + (400-540)^2 + (720-540)^2 + (500-540)^2 + (930-540)^2\right)}$$

OÙ POUVONS-NOUS DONC PLACER LA MOYENNE AVEC 95 % DE CONFIANCE?
NOUS TROUVONS NOTRE FRACTILE $t_{0,025}$ AVEC 4 DEGRÉS DE LIBERTÉ, QUI VAUT 2,78.

	$1-\alpha$	0,8	0,9	0,95	0,99
	α	0,2	0,1	0,05	0,01
	$\alpha/2$	0,1	0,05	0,025	0,005
DEGRÉS	1	3,08	6,31	12,71	63,66
DE LIBERTÉ	2	1,89	2,92	4,3	9,92
	3	1,64	2,35	3,18	5,84
	4	1,53	2,13	**2,78**	4,6
	5	1,48	2,02	2,57	4,03

ET ON LE PLONGE DANS :

$$\mu \in \bar{x} \pm 2{,}78 \frac{s}{\sqrt{n}}$$

$$\mu \in 540 \pm 2{,}78 \frac{299}{\sqrt{5}}$$

$$\mu \in \mathbf{540} \pm \mathbf{372}$$



TOUT CE QU'ON PEUT AFFIRMER AVEC 95 % DE CONFIANCE, C'EST QUE LES DOMMAGES MOYENS SERONT ENTRE 168 ET 912 €.



L'ENTREPRISE PEUT SOIT S'EN SATISFAIRE, SOIT FAIRE D'AUTRES TESTS.

POUR CALCULER CET INTERVALLE DE CONFIANCE EN UTILISANT LE *t* DE STUDENT, NOUS AVONS FAIT UNE **HYPOTHÈSE IMPLICITE.** NOUS AVONS SUPPOSÉ QUE LES COÛTS DE RÉPARATION ÉTAIENT APPROXIMATIVEMENT **DISTRIBUÉS NORMALEMENT**, AINSI SI ON CRASHE 1 000 CAMÉLÉON, L'HISTOGRAMME DES DOMMAGES DOIT ÊTRE SYMÉTRIQUE ET EN FORME DE MONTICULE. ON NE PEUT PAS DÉDUIRE CELA DE CET ÉCHANTILLON DE SEULEMENT 5 VOITURES... MAIS PEUT-ÊTRE QUE DES ANNÉES D'EXPÉRIENCE AVEC D'ANCIENS MODÈLES ONT DÉJÀ FOURNI DES HISTOGRAMMES NORMALEMENT DISTRIBUÉS DE DOMMAGE SUR LES PARTIES AVANT. CE TYPE D'INFORMATION JUSTIFIERAIT NOTRE UTILISATION DU *t* DE STUDENT.



En résumé (!), nous disposons maintenant de trois méthodes pour trouver des intervalles de confiance. Pour des proportions comme pour des moyennes avec de grands échantillons, nous cherchons $z_{\alpha/2}$ dans une table normale. Pour des moyennes de petits échantillons, disons $n \leq 30$, on trouve $t_{\alpha/2}$ dans la table de Student.



Dans tous les cas, la largeur de l'intervalle est ce fractile multiplié par l'erreur-type estimée.

$$z_{\alpha/2}s(\hat{p}) \qquad z_{\alpha/2}s(\bar{x}) \qquad t_{\alpha/2}s(\bar{x})$$

Et toutes ces erreurs-types sont proportionnelles au nombre magique.

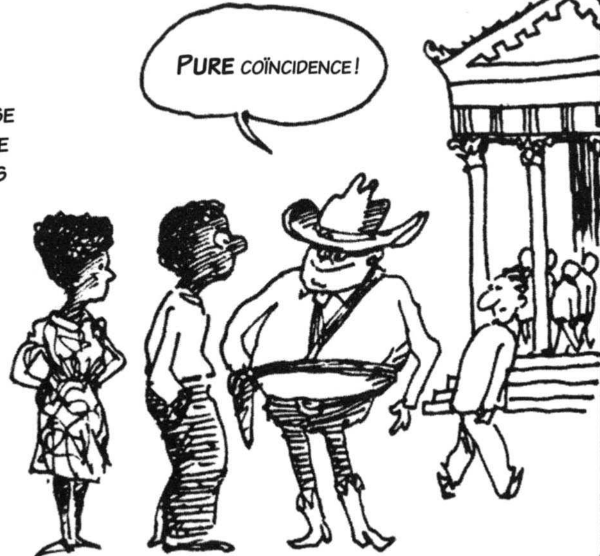
$$\frac{1}{\sqrt{n}}$$

Chapitre 8
Tests d'hypothèses

NOUS ENTRONS MAINTENANT DANS UN NOUVEAU DOMAINE. LES GOUVERNEMENTS, LES ENTREPRISES ET LES SCIENCES DURES ET MOLLES UTILISENT TOUS ET ABUSENT SOUVENT DE CES TESTS DE SIGNIFICATION. IL S'AGIT DE RÉPONDRE À LA QUESTION : **« CES OBSERVATIONS ONT-ELLES RÉELLEMENT EU LIEU PAR HASARD ? »**

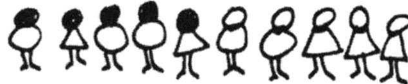


Nous commençons avec un exemple provenant du domaine juridique : plusieurs jugements ont été remis en cause dans le sud des États-Unis entre 1960 et 1980. Des témoins experts présentèrent des cas de **biais raciaux** dans la **sélection des jurys.**



Les panels de jurés sont théoriquement composés aléatoirement à partir d'une liste de citoyens éligibles. Cependant, dans les États du Sud des années 1950 et 1960, il y avait peu d'Afro-Américains dans ces jurys. Certains avocats contestèrent donc le verdict. En appel, un témoin expert en statistiques fournit cette preuve :

1) 50 % de citoyens éligibles étaient afro-américains.



2) Sur un panel de 80 jurés potentiels, seulement **quatre** étaient afro-américains.



Cela pouvait-il être le résultat d'un **pur hasard** ?

POUR SIMPLIFIER LA DISCUSSION, SUPPOSONS QUE LA SÉLECTION PARMI LES JURÉS POTENTIELS SOIT **ALÉATOIRE.** ALORS, LE NOMBRE D'AFRO-AMÉRICAINS SUR UN PANEL DE 80 PERSONNES SERA UNE **VARIABLE ALÉATOIRE X** AVEC $n = 80$ TIRAGES ET $p = 1/2$.



AINSI, LA PROBABILITÉ D'OBTENIR UN JURY DE MOINS DE QUATRE AFRO-AMÉRICAINS EST $P(X \leq 4)$, QUI S'AVÈRE ÊTRE DE L'ORDRE DE 0,0000000000000000014 (!).



COMME LA PROBABILITÉ EST INFINITÉSIMALE, LE PANEL EN QUESTION AVEC SEULEMENT QUATRE NOIRS AMÉRICAINS FOURNIT UNE **PREUVE SOLIDE** CONTRE L'**HYPOTHÈSE** D'UNE SÉLECTION ALÉATOIRE.



POUR PORTER L'ESTOCADE, LE STATISTICIEN FAIT REMARQUER QUE CETTE PROBABILITÉ EST PLUS FAIBLE QUE CELLE D'OBTENIR **TROIS QUINTES FLUSH ROYALES CONSÉCUTIVES** AU POKER.



LE JUGE **REJETTE DONC L'HYPOTHÈSE** D'UNE SÉLECTION ALÉATOIRE.



(ET IL RETIRA SA REMARQUE DE L'ENREGISTREMENT !)

Suivons le même procédé que précédemment pour développer les **quatre étapes formelles** d'un test d'hypothèse statistique.

Étape 1 : Formuler toutes les hypothèses.

H_0 **L'hypothèse nulle** est généralement que les observations sont strictement le résultat du **hasard.**

H_a **L'hypothèse alternative** indique qu'il y a un effet réel, que les observations sont le résultat de cet effet et d'une composante aléatoire.

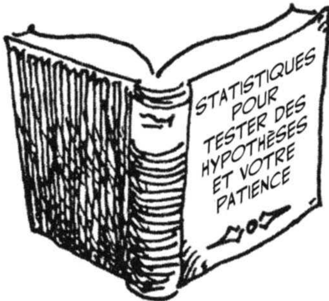


Dans l'affaire de justice, H_0 stipule que le jury était **choisi au hasard** dans la population globale. Les Afro-Américains ont une probabilité $p = 0{,}50$ d'être choisis.

H_a stipule que les Afro-Américains sont **moins sélectionnés** dans les jurys que leur proportion dans la population : $p < 0{,}50$.



Étape 2 : Le **test statistique.** Identifier la statistique qui fournira des preuves contre l'hypothèse nulle.



Ici, la statistique de test est la variable aléatoire **binomiale** X avec $p = 0{,}50$ et $n = 80$.



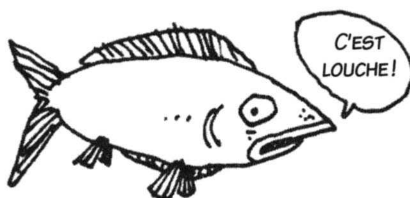
Étape 3 : *VALEUR P* : UNE CONSTATATION PROBABILISTE POUR RÉPONDRE À LA QUESTION : SI L'HYPOTHÈSE NULLE EST VRAIE, ALORS QUELLE EST LA PROBABILITÉ D'OBSERVER UNE STATISTIQUE DE TEST AU MOINS AUSSI EXTRÊME QUE CELLE OBSERVÉE ?



Étape 4 : COMPARER LA VALEUR *P* À UN **SEUIL DE SIGNIFICATION, α.**
α AGIT COMME UN SEUIL LIMITE EN DESSOUS DUQUEL NOUS NOUS ACCORDONS À DIRE QUE L'EFFET EST STATISTIQUEMENT SIGNIFICATIF. AUTREMENT DIT, SI

$$P \leq \alpha$$

ALORS NOUS **REJETONS L'HYPOTHÈSE NULLE** H_0 ET DÉCLARONS QUE QUELQUE CHOSE D'AUTRE A EU LIEU.

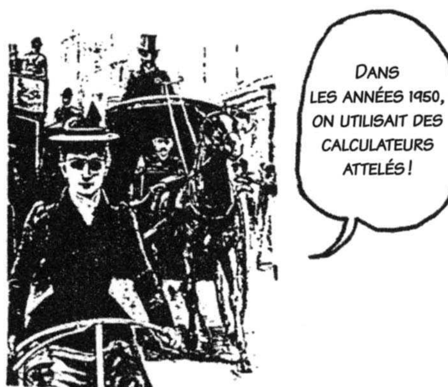


DANS NOTRE EXEMPLE, LA VALEUR *P* VAUT

$$P(X \leq 4 \mid p = 0{,}50 \text{ ET } n = 80)$$

$= 1{,}4 \times 10^{-18}$

ON A CALCULÉ LA VALEUR *P* AVEC LES MOYENS MODERNES EN UTILISANT UN LOGICIEL DE STATISTIQUES.



α EST SOUVENT PRIS À 0,05 OU 0,01. DANS LE CAS DU JURY, ET POUR L'EFFET DRAMATIQUE, LE STATISTICIEN A FAIT ALLUSION À UN α ABSURDEMENT FAIBLE DE $3{,}6 \times 10^{-18}$, QUI REPRÉSENTE LES CHANCES D'AVOIR TROIS QUINTES FLUSH ROYALES DE SUITE.





Dans les travaux scientifiques, on utilise souvent un seuil fixe de α égal à 0,05 ou 0,01. Ces valeurs sont un vestige de l'ère pré-ordinateur où l'on devait se référer à des tables imprimées pour trouver des valeurs critiques présélectionnées. Pourtant de nos jours, les journaux scientifiques publient seulement des résultats lorsque la valeur P est $\leq 0{,}05$.



Dans les cours de justice, le standard est plus flexible...



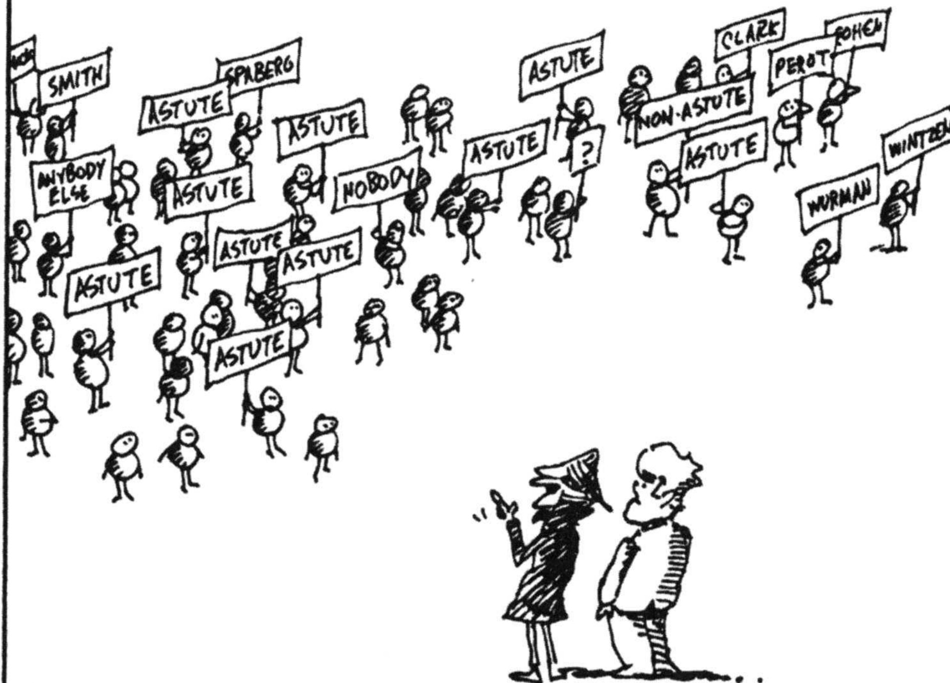
Test de signification de proportion pour de GRANDS ÉCHANTILLONS



L'EXEMPLE DU JURY EST UN CAS PARTICULIER D'UN PROBLÈME PLUS GÉNÉRAL OÙ L'HYPOTHÈSE NULLE EST DE LA FORME $p = p_0$ ET OÙ p_0 EST UNE CERTAINE PROBABILITÉ (DANS NOTRE CAS 0,5). MAINTENANT, ÉTUDIONS LE CAS GÉNÉRAL.

TESTONS L'HYPOTHÈSE $p = p_0$.

COMME D'HABITUDE, NOUS SUPPOSONS QUE NOUS AVONS UNE ÉNORME POPULATION... NOUS OBSERVONS UN GRAND ÉCHANTILLON... ET NOUS TROUVONS QU'UNE PROPORTION DE L'ÉCHANTILLON VÉRIFIE UNE PROPRIÉTÉ.



À PARTIR DE CETTE OBSERVATION, NOUS VOULONS SAVOIR SI LA VRAIE PROPORTION DE POPULATION EST (PAR EXEMPLE) PLUS GRANDE QU'UNE VALEUR p_0. PAR EXEMPLE, LE SÉNATEUR ASTUTE QUI A TROUVÉ $\hat{p} = 0{,}55$ VEUT SAVOIR SI $p > 0{,}5$, CE QUI LUI GARANTIRAIT LA MAJORITÉ.

Étape 1

L'HYPOTHÈSE NULLE EST :

$$H_0 : p = p_0$$

L'HYPOTHÈSE ALTERNATIVE DÉPEND DE LA DIRECTION DE L'EFFET CONSIDÉRÉ. DANS LE CAS DU SÉNATEUR ASTUTE :

$$H_a : p > p_0$$

MAIS DANS D'AUTRES CAS, L'HYPOTHÈSE ALTERNATIVE PEUT ÊTRE :

$$H_a : p < p_0$$

OU

$$H_a : p \neq p_0$$

PAR EXEMPLE, DANS LE CAS DE LA SÉLECTION DU JURY, L'HYPOTHÈSE ALTERNATIVE ÉTAIT :

$$H_a : p < 0{,}5$$

MAIS D'AUTRES FOIS, NOUS VOULONS SAVOIR SI P EST DIFFÉRENT OU NON D'UNE VALEUR. PAR EXEMPLE, SI ON TESTE L'HONNÊTETÉ D'UN LANCER DE PIÈCE, L'HYPOTHÈSE ALTERNATIVE EST :

$$H_a : p \neq 0{,}5$$

MAIS NOUS N'AVONS AUCUNE IDÉE A PRIORI SI LE BIAIS EST SUR LE CÔTÉ FACE OU PILE.



Étape 2

LA STATISTIQUE DE TEST EST :

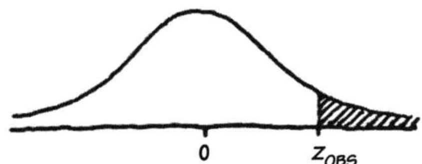
$$z_{OBS} = \frac{(\hat{p} - p_0)}{\sqrt{p_0(1 - p_0)/n}}$$

QUI MESURE L'ÉLOIGNEMENT DE $\hat{p}$ PAR RAPPORT À p_0. AVEC L'HYPOTHÈSE NULLE, z_{OBS} A UNE DISTRIBUTION NORMALE STANDARD.

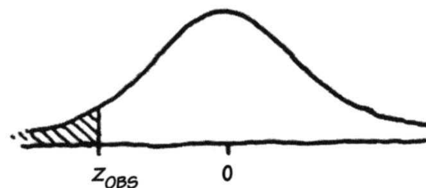
Étape 3

LA VALEUR *P* DÉPEND DU TYPE DE L'HYPOTHÈSE ALTERNATIVE :

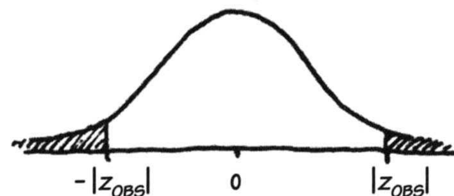
A) TEST UNILATÉRAL À DROITE, $H_a : p > p_0$ UTILISE LA VALEUR *P* DÉFINIE PAR $P(z > z_{OBS})$.



B) TEST UNILATÉRAL À GAUCHE, $H_a : p < p_0$ UTILISE LA VALEUR *P* DÉFINIE PAR $P(z < z_{OBS})$.



C) TEST BILATÉRAL À GAUCHE, $H_a : p \neq p_0$ UTILISE LA VALEUR *P* DÉFINIE PAR $P(|z| > |z_{OBS}|)$.



DANS LE CAS DU SÉNATEUR ASTUTE :

1) LES HYPOTHÈSES SONT :

$H_0 : p = 0{,}5$

$H_a : p > 0{,}5$

2) LA STATISTIQUE DE TEST EST :

$$z_{OBS} = \frac{(0{,}55 - 0{,}50)}{\sqrt{(0{,}50)(0{,}50)/1000}} = \mathbf{3{,}16}$$

3) LA VALEUR P EST :

$$P(z \geq z_{OBS}) = P(z \geq 3{,}16) = \mathbf{0{,}0008}$$

4) ÉTANT CONSERVATEUR, ASTUTE PREND UN SEUIL DE SIGNIFICATION α DE 0,01 ET IL OBSERVE QUE :

$$P(z > z_{OBS}) = 0{,}0008 < \alpha$$



LE SÉNATEUR REJETTE DONC L'HYPOTHÈSE NULLE, ET LUI ET SON ÉQUIPE PEUVENT ÊTRE SÛRS D'ÊTRE EN TÊTE.



NDT : LORSQUE L'HYPOTHÈSE ALTERNATIVE EST UNE INÉGALITÉ (DU TYPE $H_0 : p < p_0$), CERTAINS STATISTICIENS PRENNENT COMME HYPOTHÈSE NULLE LE CONTRAIRE DE L'HYPOTHÈSE ALTERNATIVE (SOIT $H_0 : p \geq p_0$). CELA NE CHANGE EN RIEN L'ANALYSE CAR C'EST L'HYPOTHÈSE ALTERNATIVE QUI DÉFINIT LE CRITÈRE DE REJET OU NON.

Test de **MOYENNE** de population pour de **GRANDS ÉCHANTILLONS**



MAIS BIEN SÛR L'ÉPICERIE N'A AUCUNE INTENTION DE PESER CHACUNE DES BOÎTES EXPÉDIÉES. SES EMPLOYÉS SE CONTENTERONT D'UTILISER LES STATISTIQUES !



TOUT D'ABORD, ILS CHOISISSENT LEURS HYPOTHÈSES :

$$H_0 : \mu = 16 \text{ OZ.}$$

$$H_a : \mu < 16 \text{ OZ.}$$

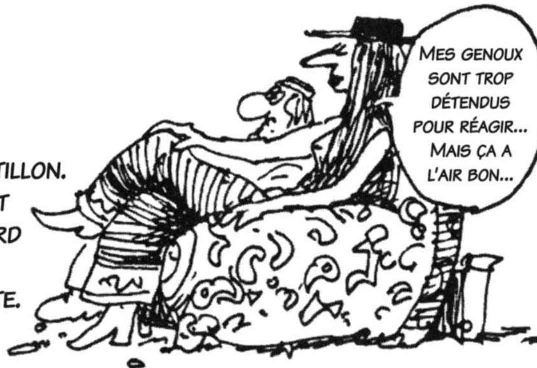
REJETER L'HYPOTHÈSE IMPLIQUE DE REFUSER LA LIVRAISON DES CÉRÉALES.



ENSUITE, ILS CHOISISSENT UNE STATISTIQUE DE TEST. MAINTENANT, CELA DEVRAIT ÊTRE COMME UN RÉFLEXE D'EXTENSION DU GENOU DE DIRE QUE L'ÉCART DE L'ÉCHANTILLON À LA MOYENNE EST :

$$\frac{\bar{X} - \mu_0}{s(\bar{X})} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

OÙ s EST L'ÉCART-TYPE DE L'ÉCHANTILLON. SOUS L'HYPOTHÈSE NULLE, CELA SUIT PRESQUE UNE LOI NORMALE STANDARD AVEC DE GRANDS ÉCHANTILLONS, D'APRÈS LE THÉORÈME CENTRAL LIMITE.



EN SAUTANT L'ÉTAPE 3 POUR LE MOMENT, ILS FIXENT UN SEUIL DE SIGNIFICATION. AYANT TOUS RATÉ LA PLACE DE MAJOR EN SCIENCES, NOS ÉPICIERS PENSENT QUE α = 0,05 SONNE BIEN.



AU MÊME MOMENT UN CAMION CHARGÉ DE 10 000 BOÎTES DE CÉRÉALES ARRIVE À LA PORTE.

ILS PRENNENT UN ÉCHANTILLON ALÉATOIRE SIMPLE DE 49 BOÎTES, LES PÈSENT ET DÉTERMINENT LES STATISTIQUES DE L'ÉCHANTILLON.

$\bar{x}$ = **15,9** OZ
s = **0,35** OZ

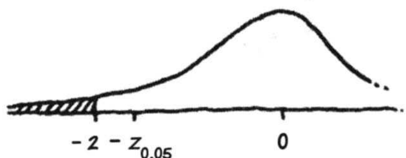
C'EST UN PEU LÉGER, MAIS EST-CE SIGNIFICATIF ?



ILS RENTRENT LES VALEURS DANS LA STATISTIQUE DE TEST POUR TROUVER :

$$z_{OBS} = \frac{15{,}9 - 16}{0{,}35/\sqrt{49}} = \mathbf{-2}$$

MAINTENANT, ILS CALCULENT LEUR VALEUR P : $P(z < -2 \mid H_0) = \mathbf{0{,}0227}$



COMME C'EST MOINS QUE LE SEUIL DE SIGNIFICATION DE 0,05, L'ÉPICERIE AUTHENTIQUE REJETTE L'HYPOTHÈSE NULLE, ET LA MARCHANDISE.





Test de **MOYENNE** de population pour de **PETITS ÉCHANTILLONS**



NOUS REVENONS DANS L'ENTREPRISE CAMÉLÉON AUTOMOBILES ET SES CRASH-TESTS. LA **COMPAGNIE D'ASSURANCES INTÈGRE** ASSURERA UNE AUTO SEULEMENT SI LA MOYENNE DES COÛTS DE RÉPARATION APRÈS UNE COLLISION À 16 km/h S'ÉLÈVE À MOINS DE 1000 €. LA COMPAGNIE UTILISE UN SEUIL DE SIGNIFICATION STANDARD $\alpha = 0{,}05$. AINSI :

$H_0 : \mu \geq 1000$ € LE COÛT MOYEN EST TROP ÉLEVÉ.

$H_a : \mu < 1000$ € LE COÛT MOYEN EST CORRECT.

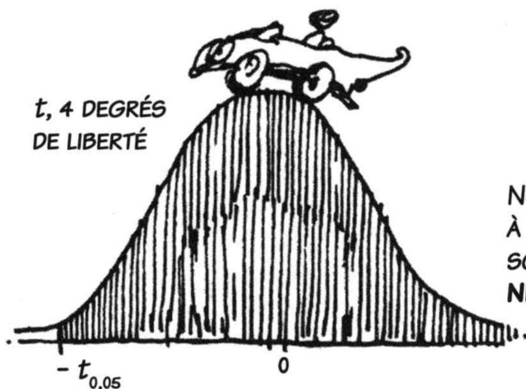


LA STATISTIQUE DE TEST EST :

$$t = \frac{\bar{X} - \mu_0}{s(\bar{X})}$$

OÙ μ_0 EST LA MOYENNE HYPOTHÉTIQUE DE 1000 €.

SOUS L'HYPOTHÈSE NULLE, LA STATISTIQUE A UNE DISTRIBUTION t DE STUDENT AVEC 4 DEGRÉS DE LIBERTÉ.



NOUS VOULONS QUE LE t OBSERVÉ SOIT À GAUCHE DE $-t_{0,05}$ (CAR DE FAIBLES $\bar{x}$ SONT SOUHAITÉS, $\bar{x} - \mu_0$ DOIT ÊTRE **NÉGATIF** POUR CONFIRMER H_a).

		α	
Degrés de liberté	0,05	0,025	0,005
1	6,31	12,71	63,66
2	2,92	4,30	9,92
3	2,35	3,18	5,84
4	**2,13**	2,78	4,6
5	2,01	2,57	4,03

DANS LA TABLE DES VALEURS CRITIQUES t, ON VOIT QUE $t_{0,05}$ = **2,13**, ON REJETTE DONC H_0 SI :

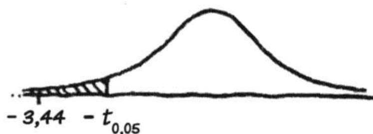
$$t_{OBS} \leq -t_{0,05} = -2,13$$

NOUS SAVONS (VOIR CHAPITRE 7) QUE $\bar{x}$ = 540 € ET s = 299 € POUR NOTRE PETIT ÉCHANTILLON DE 5 VOITURES, AINSI ON OBTIENT :

$$t_{OBS} = \frac{540 - 1000}{299/\sqrt{5}}$$

$$= -3,44 < -t_{0,05}$$





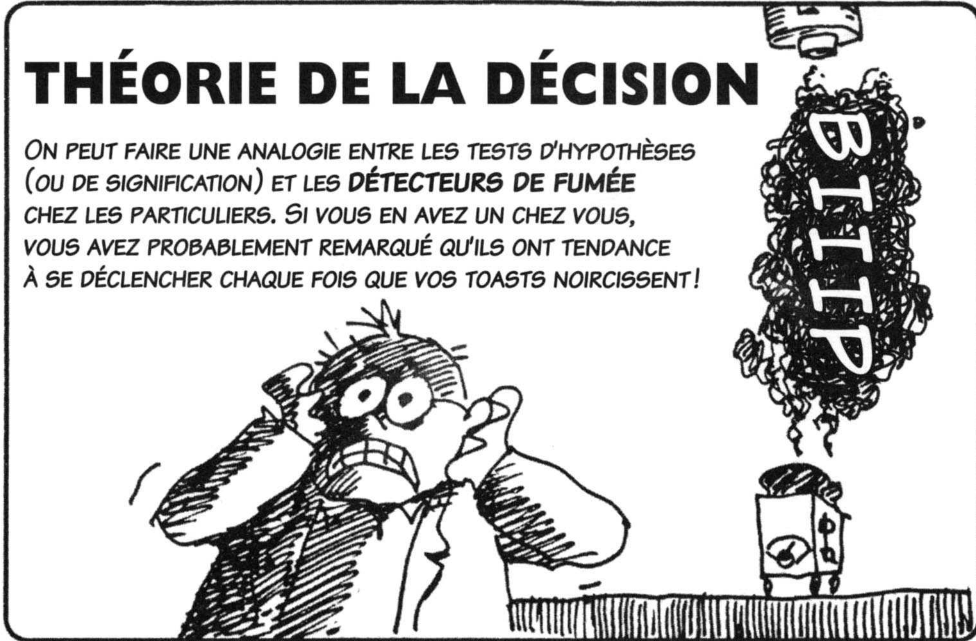
LA VOITURE PASSE LE TEST... H_0 EST REJETÉE... ET LA POLICE D'ASSURANCE EST ÉMISE.



C'EST UN EXEMPLE D'**ÉCHANTILLONNAGE D'ACCEPTATION.** L'HYPOTHÈSE NULLE EST QUE LES COÛTS DE RÉPARATION NE SONT PAS ACCEPTABLES. L'ENTREPRISE EST PRÉSUMÉE COUPABLE ET DOIT FOURNIR DES PREUVES SUFFISANTES DE SON INNOCENCE EN MONTRANT QUE LE PRODUIT EST CONFORME AUX SPÉCIFICATIONS.

THÉORIE DE LA DÉCISION

On peut faire une analogie entre les tests d'hypothèses (ou de signification) et les **détecteurs de fumée** chez les particuliers. Si vous en avez un chez vous, vous avez probablement remarqué qu'ils ont tendance à se déclencher chaque fois que vos toasts noircissent !



C'est ce qui s'appelle une **erreur de type I :** une alarme sans feu. Inversement, une **erreur de type II** est un feu sans alarme. Tout cuisinier sait éviter les erreurs de type I : il suffit d'**enlever les piles.** Malheureusement, cela augmente la fréquence des erreurs de type II !



De façon similaire, réduire les risques d'erreurs de type II, en rendant, par exemple, l'alarme hypersensible, peut accroître le nombre de fausses alarmes.

ON PEUT RÉSUMER CES RÉSULTATS DANS UNE **TABLE DE DÉCISION :**

	PAS DE FEU	FEU
PAS D'ALARME	PAS D'ERREUR	TYPE II
ALARME	TYPE I	PAS D'ERREUR

MAINTENANT, VOYONS L'HYPOTHÈSE NULLE COMME ÉTANT LA CONDITION QU'IL N'Y A **PAS DE FEU,** ALORS QUE L'HYPOTHÈSE ALTERNATIVE SIGNIFIE QU'IL Y A LE FEU. L'ALARME CORRESPOND AU REJET DE L'HYPOTHÈSE NULLE.

	VÉRITABLE ÉTAT H_0	H_a
ACCEPTER H_0	PAS D'ERREUR	TYPE II
REJETER H_0	TYPE I	PAS D'ERREUR

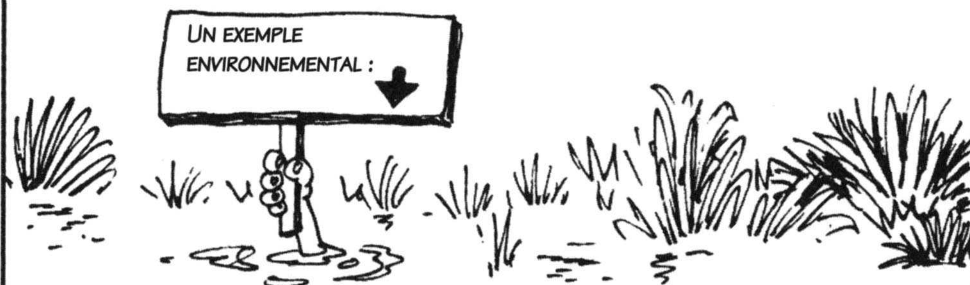
TOUS LES TESTS DE SIGNIFICATION QUE NOUS VENONS D'EFFECTUER DANS CE CHAPITRE METTENT L'ACCENT SUR LA PROBABILITÉ DE COMMETTRE UNE ERREUR DE TYPE I (C'EST-À-DIRE LA PROBABILITÉ QUE NOS OBSERVATIONS SE RÉALISENT LORSQUE H_0 EST VRAIE). NOUS AVIONS POSÉ :

$$P(\text{REJETER } H_0 \mid H_0) = P(\text{ERREUR DE TYPE I} \mid H_0) = \alpha$$

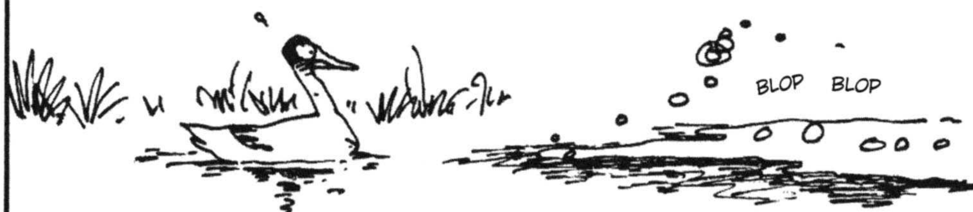
$1 - \alpha$ MESURE NOTRE CONFIANCE QUE QUAND L'ALARME SONNE, IL Y A EFFECTIVEMENT LE FEU. UNE CONFIANCE ÉLEVÉE SIGNIFIE QUE LE DÉCLENCHEMENT DE FAUSSES ALARMES EST RARE.



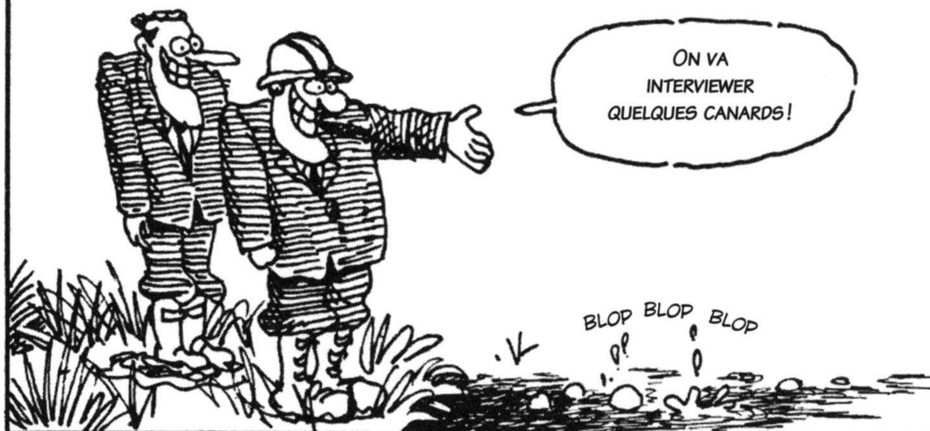
MAIS PARFOIS, CE QUE NOUS VOULONS VRAIMENT CONNAÎTRE C'EST LA PROBABILITÉ DE FAIRE UNE **ERREUR DE TYPE II.** EN D'AUTRES TERMES, QUELLE EST LA SENSIBILITÉ DE NOTRE « SYSTÈME D'ALARME » QUAND L'HYPOTHÈSE ALTERNATIVE EST VÉRIFIÉE ?



DANS LE PASSÉ, LES USINES QUI DÉVERSAIENT DES PRODUITS CHIMIQUES DANS LES COURS D'EAU DEVAIENT PROUVER QUE CETTE ACTION N'AVAIT AUCUN IMPACT SUR LA VIE AQUATIQUE EN AVAL. C'EST H_0. LE POLLUEUR POUVAIT CONTINUER TANT QUE L'HYPOTHÈSE NULLE N'ÉTAIT PAS REJETÉE AU SEUIL DE SIGNIFICATION DE 5 %.



AINSI LE POLLUEUR QUI PENSAIT ÊTRE EN VIOLATION DES NORMES CONCEVAIT LUI-MÊME UN PROGRAMME DE CONTRÔLE DE POLLUTION INEFFICACE.



LE POLLUEUR EST RAVI PUISQUE, COMME NOTRE ALARME INCENDIE SANS PILES, SON TEST A TRÈS PEU OU AUCUNE CHANCE DE DÉCLENCHER L'ALARME.



FORMALISONS CETTE IDÉE. POUR DÉCRIRE LA **PROBABILITÉ D'UNE ERREUR DE TYPE II,** NOUS SORTONS UNE AUTRE LETTRE GRECQUE, BÊTA OU β.

$$\beta = P(\text{ACCEPTER } H_0 \mid H_a)$$
$$= P(\text{ERREUR DE TYPE II} \mid H_a)$$

LA **PUISSANCE** D'UN TEST EST DÉFINIE PAR $1 - \beta$. IL S'AGIT DE :

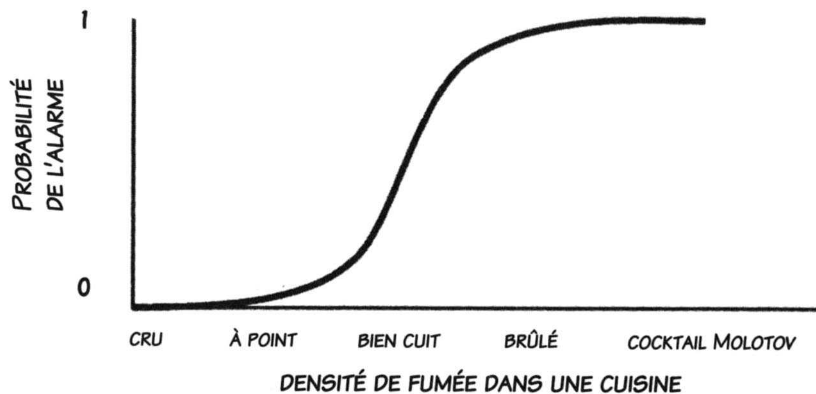
$$P(\text{REJETER } H_0 \mid H_a).$$



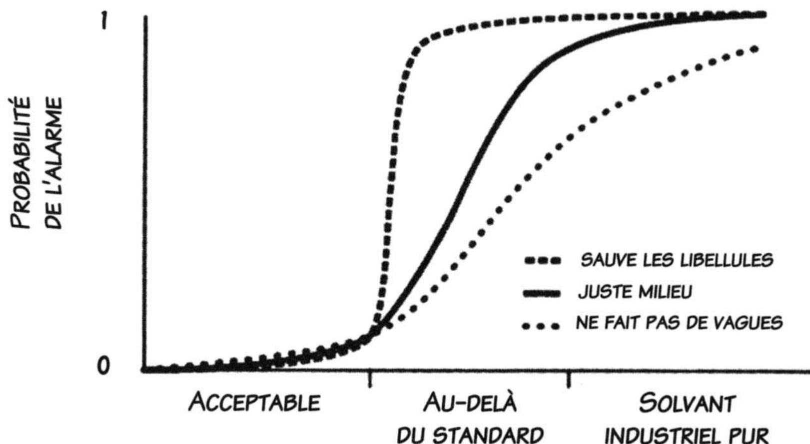
VOUS SEREZ CONTENT D'APPRENDRE QUE LES AGENCES ENVIRONNEMENTALES ONT ÉVOLUÉ VERS DES PROGRAMMES DE CONTRÔLE DE POLLUTION QUI DOIVENT MONTRER UNE PROBABILITÉ ÉLEVÉE DE DÉTECTER LES ÉVÉNEMENTS POLLUANTS IMPORTANTS. L'**ANALYSE DE PUISSANCE** REQUISE RÉVÈLE SOUVENT LES FAILLES CACHÉES DES PROGRAMMES DE CONTRÔLES.



UNE FAÇON DE VISUALISER L'EFFET DE LA PUISSANCE D'UN TEST EST DE TRACER LA PROBABILI DE REJETER H_0 COMME UNE FONCTION DE L'ÉTAT DU SYSTÈME. DANS LE CAS DE L'ALARME INCENDIE, LA PROBABILITÉ CROÎT VERS 1 QUAND LA FUMÉE S'ÉPAISSIT.



POUR L'EXEMPLE SUR LA QUALITÉ DE L'EAU, L'AXE HORIZONTAL EST LA VRAIE CONCENTRATION DE POLLUANTS DANS L'EAU.



VOICI LES COURBES DE PUISSANCE DE TROIS PROGRAMMES DE CONTRÔLE. LE « **SAUVE LES LIBELLULES** » (COÛTE 5 MILLIONS DE €), LE « **JUSTE MILIEU** » (COÛTE 500 000 €), LE « **NE FAIT PAS DE VAGUES** » (COÛTE AUSSI 500 000 €, MAIS IL GARANTIT UN BON SPECTACLE !). PLUS LA PUISSANCE DU TEST EST IMPORTANTE, PLUS LA COURBE EST ABRUPTE.



POURQUOI AVEZ-VOUS ALORS CETTE **SENSATION DE VIDE** DANS VOTRE ESTOMAC ? C'EST PARCE QUE POUR UTILISER CES IDÉES DE FAÇON PRATIQUE, NOUS DEVONS POUVOIR LES APPLIQUER À UNE VARIÉTÉ DE SITUATIONS QUE NOUS N'AVONS PAS ENCORE VUES. C'EST PAR LÀ QUE NOUS ALLONS POURSUIVRE AVEC LA **COMPARAISON DE DEUX POPULATIONS.**



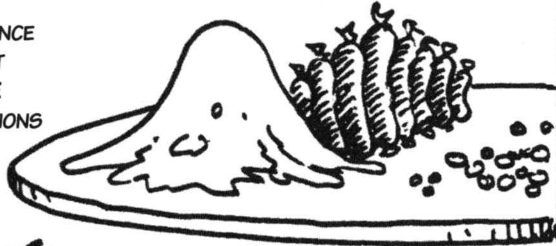


Chapitre 9
Comparaison de deux populations

DANS LEQUEL NOUS APPRENONS DE NOUVELLES RECETTES EN UTILISANT DE VIEUX INGRÉDIENTS.



LES DEUX DERNIERS CHAPITRES NOUS ONT EXPLIQUÉ LES INTERVALLES DE CONFIANCE ET LES TESTS D'HYPOTHÈSES EN UTILISANT LA **VIANDE** ET LES **POMMES DE TERRE** DES MODÈLES ALÉATOIRES : LES DISTRIBUTIONS NORMALES ET BINOMIALES.

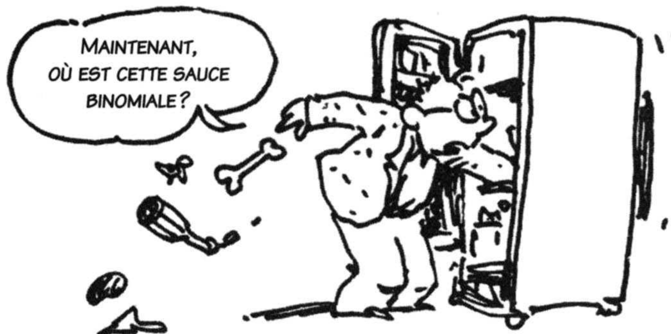


AVEC LA LOI NORMALE DANS LE RÔLE DES PATATES !

MAIS CE QUI REND LES STATISTIQUES AUSSI STIMULANTES QUE LA CUISINE, C'EST LA DIVERSITÉ. COMME UN EXPERT EN CUISINE, LE STATISTICIEN PEUT « GOÛTER » LES INGRÉDIENTS D'UN PROBLÈME ET TROUVER LA FAÇON LA PLUS EFFICACE DE LES COMBINER DANS UNE RECETTE STATISTIQUE.



(LA RAISON POUR LAQUELLE LES LIVRES DE CUISINE ET LES MANUELS DE STATISTIQUES SONT SI LOURDS EST QU'ILS FOURNISSENT TOUS LES DEUX DES SOLUTIONS À UNE GRANDE VARIÉTÉ DE SITUATIONS !)



DANS CE CHAPITRE, NOUS ALLONS UTILISER NOS MÉTHODES « VIANDE-ET-PATATE » DANS DE NOUVELLES RECETTES QUI VONT NOUS AIDER À RÉPONDRE AUX QUESTIONS SUIVANTES :



EST-CE QUE PRENDRE DE L'ASPIRINE RÉGULIÈREMENT RÉDUIT LE RISQUE D'INFARCTUS ?



EST-CE QU'UN PESTICIDE PARTICULIER AUGMENTE LE RENDEMENT AGRICOLE ?



EST-CE QUE LES HOMMES ET LES FEMMES FAISANT LE MÊME TRAVAIL ONT DES SALAIRES DIFFÉRENTS ?



L'INGRÉDIENT COMMUN À CES QUESTIONS EST QU'IL EST POSSIBLE D'Y RÉPONDRE **EN COMPARANT DEUX ÉCHANTILLONS ALÉATOIRES INDÉPENDANTS,** UN POUR CHAQUE POPULATION.

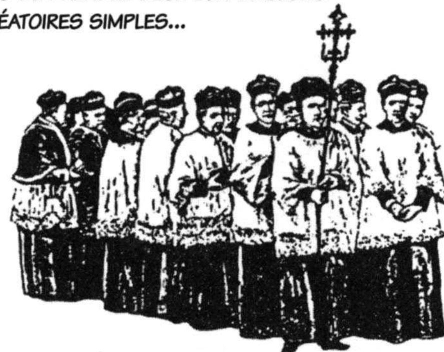


PESTICIDE



SANS PESTICIDE

ET À LA FIN DU CHAPITRE, NOUS VERRONS UNE AUTRE MÉTHODE QUI PERMET DE COMPARER DEUX MOYENNES ET QUI NE NÉCESSITE PAS DE PRENDRE DEUX ÉCHANTILLONS ALÉATOIRES SIMPLES...



Comparaison de TAUX DE SUCCÈS (ou d'échecs) de deux populations

NOUS COMMENÇONS PAR UNE EXPÉRIENCE, EN PARTIE MENÉE PAR UNE ÉTUDE DE HARVARD, QUI CHERCHAIT À DÉTERMINER L'EFFICACITÉ DE L'**ASPIRINE** POUR RÉDUIRE LES CRISES CARDIAQUES. COMME DANS LA PLUPART DES ESSAIS CLINIQUES, LES RISQUES QU'UN INDIVIDU AIT LA MALADIE – ICI UNE CRISE CARDIAQUE – SONT TRÈS MINCES SUR UNE ANNÉE. MAIS NOUS VOULONS UNE RÉPONSE RAPIDE ! COMMENT FAIRE ?



LA SOLUTION SIMPLE MAIS COÛTEUSE EST DE TESTER UN GRAND NOMBRE D'INDIVIDUS EN PEU DE TEMPS. DANS CETTE ÉTUDE, 22 071 SUJETS (TOUS DES DOCTEURS VOLONTAIRES) FURENT ALÉATOIREMENT RÉPARTIS EN **DEUX GROUPES.**



LE GROUPE 1 PRENAIT UN **PLACEBO,** UNE PILULE IDENTIQUE À L'ASPIRINE MAIS SANS ASPIRINE.



LE GROUPE 2 RECEVAIT UNE ASPIRINE PAR JOUR.

SUR UNE PÉRIODE AVOISINANT LES CINQ ANS*, LES ENQUÊTEURS ENREGISTRÈRENT LES RÉPONSES : CRISE CARDIAQUE OU NON. LE RÉSULTAT (DANS LES NOMBRES QUI SUIVENT NOUS AVONS SOMMÉ LES CRISES CARDIAQUES FATALES OU NON) :

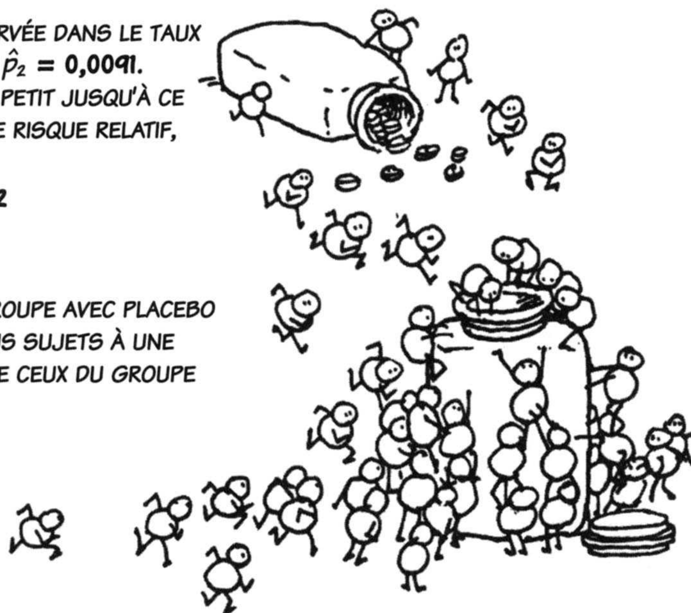


	ATTAQUES	PAS D'ATTAQUES	n	TAUX D'ATTAQUES
PLACEBO	**239**	**10 795**	**11 034**	$\hat{p}_1 = \frac{239}{11\,034} = 0{,}0217$
ASPIRINE	**139**	**10 898**	**11 037**	$\hat{p}_2 = \frac{139}{11\,037} = 0{,}0126$

LA DIFFÉRENCE OBSERVÉE DANS LE TAUX DE SUCCÈS EST $\hat{p}_1 - \hat{p}_2 = 0{,}0091$. CELA PEUT PARAÎTRE PETIT JUSQU'À CE QUE L'ON REGARDE LE RISQUE RELATIF,

$$\frac{\hat{p}_1}{\hat{p}_2} = \frac{0{,}0217}{0{,}0126} = 1{,}72$$

LES MEMBRES DU GROUPE AVEC PLACEBO ÉTAIENT 1,72 FOIS PLUS SUJETS À UNE CRISE CARDIAQUE QUE CEUX DU GROUPE AVEC L'ASPIRINE.



*L'ÉTUDE FUT ARRÊTÉE AVANT SON TERME À CAUSE DE SON RÉSULTAT POSITIF. IL AURAIT ÉTÉ IMPRUDENT ET IRRÉALISTE DE CACHER LES RÉSULTATS AU GROUPE QUI PRENAIT LE PLACEBO.

Le modèle : LES OBSERVATIONS DES GROUPES « PLACEBO ET ASPIRINE » SONT DES ÉCHANTILLONS INDÉPENDANTS DE DEUX POPULATIONS BINOMIALES. PAR SOUCI DE COHÉRENCE, NOUS POSONS QU'UNE CRISE CARDIAQUE DÉFINIT UN **SUCCÈS** (!).





POPULATION 1 PLACEBO
PROBABILITÉ DE SUCCÈS = p_1

POPULATION 2 ASPIRINE
PROBABILITÉ DE SUCCÈS = p_2

L'OBJECTIF EST D'ESTIMER LA VRAIE DIFFÉRENCE $p_1 - p_2$.

POUR CHAQUE POPULATION (EN FAIT POUR DE GRANDS ÉCHANTILLONS DE LA POPULATION GÉNÉRALE), NOUS AVONS LES VARIABLES ALÉATOIRES USUELLES :

X_1 NOMBRE DE SUCCÈS DE L'ÉCHANTILLON 1

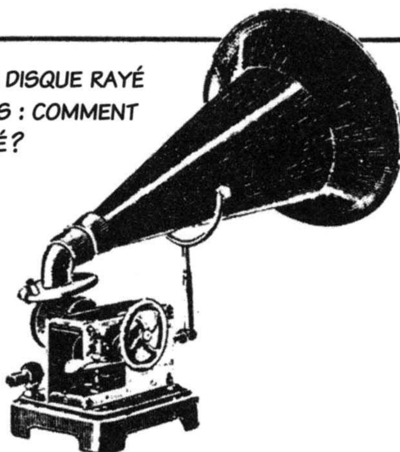
X_2 NOMBRE DE SUCCÈS DE L'ÉCHANTILLON 2

$\hat{p}_1 = \frac{X_1}{n_1}$ PROPORTION DE SUCCÈS DE L'ÉCHANTILLON 1

$\hat{p}_2 = \frac{X_2}{n_2}$ PROPORTION DE SUCCÈS DE L'ÉCHANTILLON 2

ET UN ESTIMATEUR DU TAUX DE DIFFÉRENCE QUI EST $\hat{P}_1 - \hat{P}_2$.

ET MAINTENANT, TEL UN DISQUE RAYÉ NOUS NOUS DEMANDONS : COMMENT $\hat{P}_1 - \hat{P}_2$ EST-IL DISTRIBUÉ ?



Distribution d'échantillonnage de $\hat{P}_1 - \hat{P}_2$

POUR DE GRANDS ÉCHANTILLONS, LA DISTRIBUTION DE $\hat{P}_1 - \hat{P}_2$ EST APPROXIMATIVEMENT NORMALE COMME DANS LE CAS D'UN UNIQUE ÉCHANTILLON. ON PEUT TRANSFORMER EN z POUR OBTENIR (APPROXIMATIVEMENT) UNE LOI NORMALE STANDARD

$$z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sigma(\hat{P}_1 - \hat{P}_2)}$$



MAIS COMMENT TROUVER L'ÉCART-TYPE DU DÉNOMINATEUR?

COMME LES DEUX ÉCHANTILLONS SONT INDÉPENDANTS, LES VARIABLES ALÉATOIRES $\hat{P}_1$ ET $\hat{P}_2$ LE SONT AUSSI ET LEURS VARIANCES **S'ADDITIONNENT.**

$$\sigma^2(\hat{P}_1 - \hat{P}_2) = \sigma^2(\hat{P}_1) + \sigma^2(\hat{P}_2)$$

DONC

$$\sigma(\hat{P}_1 - \hat{P}_2) = \sqrt{\sigma^2(\hat{P}_1) + \sigma^2(\hat{P}_2)}$$

MAINTENANT QUE NOUS CONNAISSONS LA DISTRIBUTION DE LA STATISTIQUE DE TEST, ON PEUT ESTIMER DES **INTERVALLES DE CONFIANCE** ET **TESTER L'HYPOTHÈSE** QUE L'ASPIRINE RÉDUIT LE RISQUE DE CRISE CARDIAQUE.



Intervalle de confiance de $p_1 - p_2$

COMME D'HABITUDE, L'INTERVALLE DE CONFIANCE POUR NOTRE ESTIMATION EST DU TYPE :

$$p_1 - p_2 \in \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} s(\hat{p}_1 - \hat{p}_2)$$

- $p_1 - p_2$: VRAIE DIFFÉRENCE DES PROPORTIONS DE POPULATION
- $\hat{p}_1 - \hat{p}_2$: DIFFÉRENCE OBSERVÉE
- $z_{\alpha/2}$: FRACTILE OU VALEUR CRITIQUE
- $s(\hat{p}_1 - \hat{p}_2)$: ERREUR-TYPE ESTIMÉE

LES VARIANCES DE $\hat{P}_1$ ET $\hat{P}_2$ S'AJOUTENT, DONC L'ERREUR-TYPE DEVIENT :

$$s(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

DANS L'ÉTUDE SUR L'ASPIRINE, L'ERREUR-TYPE EST :

$$\sqrt{\frac{(0{,}0217)(0{,}9783)}{11\,034} + \frac{(0{,}0126)(0{,}9874)}{11\,037}}$$

$$= 0{,}00175$$



DANS LE CAS DE L'ASPIRINE, POUR OBTENIR UN INTERVALLE DE CONFIANCE À 95 %, ON REMPLACE PAR LES VALEURS :

$$p_1 - p_2 \in 0{,}0091 \pm (1{,}96)(0{,}00175)$$

$$p_1 - p_2 \in \mathbf{0{,}0091 \pm 0{,}0034}$$



NOUS SOMMES CONFIANTS AU MOINS À 95 % QUE LA DIFFÉRENCE DANS LE TAUX DE CRISES CARDIAQUES EST ENTRE **0,0057** ET **0,0125,** UN NOMBRE DÉFINITIVEMENT POSITIF ! NOUS SOMMES DONC SÛRS AU MOINS À 95 % QUE L'ASPIRINE ABAISSE VRAIMENT LE TAUX DE CRISES CARDIAQUES.



Tests d'hypothèses

LA QUESTION FORMELLE D'UN TEST D'HYPOTHÈSE EST :



H_0 : L'HYPOTHÈSE NULLE EST QUE L'ASPIRINE N'A PAS D'EFFET : $p_1 = p_2$.

H_a : L'ALTERNATIVE EST QUE L'ASPIRINE RÉDUIT LE TAUX DE CRISES CARDIAQUES : $p_1 > p_2$.

MAINTENANT, IL NOUS FAUT UNE STATISTIQUE DISTRIBUÉE NORMALEMENT POUR LAQUELLE H_0 EST VRAIE.



NOTEZ QUE SOUS H_0, LES DEUX PROPORTIONS SONT ÉGALES $p_1 = p_2 = p$... ON MÉLANGE LES DONNÉES POUR AVOIR LA PROPORTION DE CRISES CARDIAQUES DANS LES DEUX ÉCHANTILLONS PRIS ENSEMBLE :

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

QUAND L'HYPOTHÈSE NULLE EST VRAIE, L'ERREUR-TYPE NOTÉE s_0 (INDICE 0 POUR H_0 VRAIE) DÉPEND UNIQUEMENT DE L'ESTIMATION **MUTUALISÉE :**

$$s_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

ET ON A CETTE STATISTIQUE DE TEST :

$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_0(\hat{p}_1 - \hat{p}_2)}$$

(LE NUMÉRATEUR EST EN GÉNÉRAL PLUTÔT $\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)$, MAIS H_0 SUPPOSE QUE $p_1 - p_2 = 0$.)



POUR L'ÉTUDE DES ASPIRINES, ON TROUVE :

$$\hat{p} = \frac{378}{22\,071} = 0{,}0171 \text{ ET } 1 - \hat{p} = 0{,}9829$$

$s_0(\hat{p}_1 - \hat{p}_2) = 0{,}00175$ DONC :

$$z_{OBS} = \frac{\hat{p}_1 - \hat{p}_2}{s_0(\hat{p}_1 - \hat{p}_2)} = \frac{0{,}0091}{0{,}00175}$$

$$= 5{,}20$$

z_{OBS} EST À PLUS DE **CINQ ÉCARTS-TYPES** DE ZÉRO, L'EFFET EST FORTEMENT POSITIF. EN UTILISANT UNE TABLE OU UN ORDINATEUR, ON OBTIENT LA VALEUR P :

VALEUR DE $P = P(z \geq z_{OBS}) = P(z \geq \mathbf{5{,}2}) = \mathbf{0{,}0000001}$



SI L'HYPOTHÈSE NULLE ÉTAIT VRAIE, LA PROBABILITÉ D'OBSERVER UN EFFET AUSSI LARGE EST DE UN SUR **DIX MILLIONS** – UNE PREUVE SOLIDE **CONTRE** H_0 !

La recette générale :



POUR TESTER L'HYPOTHÈSE NULLE :

$$H_0 : p_1 = p_2$$

ON CALCULE LA STATISTIQUE DE TEST :

$$z_{OBS} = \frac{\hat{p}_1 - \hat{p}_2}{s_0(\hat{P}_1 - \hat{P}_2)}$$

OÙ s_0 EST CALCULÉE EN UTILISANT LA PROBABILITÉ MUTUALISÉE EN MÉLANGEANT LES GROUPES.



LA VALEUR P PERTINENTE DÉPEND DE L'HYPOTHÈSE ALTERNATIVE :

A) BILATÉRAL, $H_a : p_1 \neq p_2$



VALEUR DE $P = P(|z| > z_{OBS})$

B) UNILATÉRAL DROIT, $H_a : p_1 > p_2$



VALEUR DE $P = P(z > z_{OBS})$

C) UNILATÉRAL GAUCHE, $H_a : p_1 < p_2$



VALEUR DE $P = P(z < z_{OBS})$

L'ANALYSE DE L'ÉTUDE SUR L'ASPIRINE DÉPENDAIT DE CERTAINES CARACTÉRISTIQUES DE L'EXPÉRIENCE CONÇUES POUR GARANTIR LE HASARD ET ÉLIMINER LES BIAIS.







LES POINTS 1 ET 2 SONT ESSENTIELS DANS LA CONCEPTION DE LA PLUPART DES ESSAIS CLINIQUES. MAIS LE POINT 3 N'EST PAS INDISPENSABLE. IL EXISTE DE BONS TESTS SUR DE PETITS ÉCHANTILLONS, DISPONIBLES DANS DES LOGICIELS DE STATISTIQUES. CES PROCÉDURES **NON PARAMÉTRIQUES** DÉPENDENT DE CALCULS DE PROBABILITÉS SIMPLES MAIS LONGS DU TYPE DE CEUX QUE NOUS AVONS RENCONTRÉS DANS LE CHAPITRE 4.

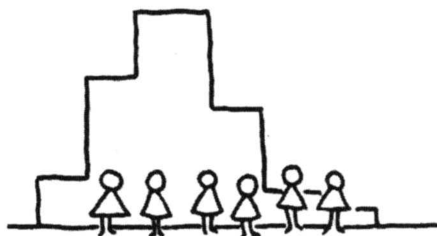


Comparaison de MOYENNES de deux populations

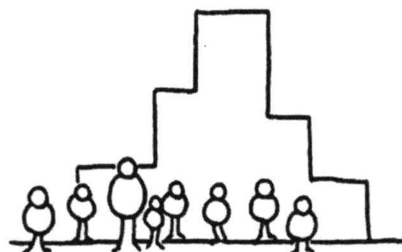
SUPPOSONS QUE L'ON VEUILLE COMPARER LE SALAIRE MOYEN DES HOMMES À CELUI DES FEMMES EMPLOYÉES POUR LE MÊME TRAVAIL DANS UNE ENTREPRISE.



LA POPULATION 1 EST CELLE DES FEMMES, ET LA POPULATION 2 CELLE DES HOMMES.



LA POPULATION 1 A UN SALAIRE MOYEN μ_1 ET UN ÉCART-TYPE σ_1.



LA POPULATION 2 A UN SALAIRE MOYEN μ_2 ET UN ÉCART-TYPE σ_2.

UN ÉCHANTILLON ALÉATOIRE SIMPLE DE TAILLE n_1 POUR LE GROUPE 1 ET DE TAILLE n_2 POUR LE GROUPE 2 FOURNIT DES MOYENNES D'ÉCHANTILLONS DE $\bar{x}_1$ ET $\bar{x}_2$ AVEC DES ÉCARTS-TYPES RESPECTIVEMENT DE s_1 ET s_2. L'ESTIMATEUR DE $\mu_1 - \mu_2$ EST :

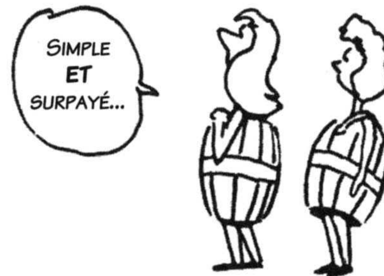
$$\bar{X}_1 - \bar{X}_2$$

À QUEL POINT NOTRE ESTIMATEUR $\bar{X}_1 - \bar{X}_2$ EST-IL BON ?
POUR DE GRANDS ÉCHANTILLONS, C'EST PRESQUE NORMAL D'APRÈS LE THÉORÈME CENTRAL LIMITE ET L'ERREUR-TYPE VAUT

$$s(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



(LES VARIANCES S'AJOUTENT CAR LES ÉCHANTILLONS SONT INDÉPENDANTS.)
MAINTENANT NOUS POUVONS DIRECTEMENT POURSUIVRE AVEC LES **intervalles de confiance :** POUR DE GRANDS ÉCHANTILLONS, L'INTERVALLE AVEC $(1 - \alpha)$ DE CONFIANCE POUR LA DIFFÉRENCE DE MOYENNE EST :



$$\mu_1 - \mu_2 \in \bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} s(\bar{X}_1 - \bar{X}_2)$$

Test d'hypothèse : ON POSE L'HYPOTHÈSE NULLE QUI DIT QUE LES MOYENNES DE POPULATION SONT ÉGALES.

$$H_0 : \mu_1 = \mu_2$$

LA STATISTIQUE DE TEST VAUT :

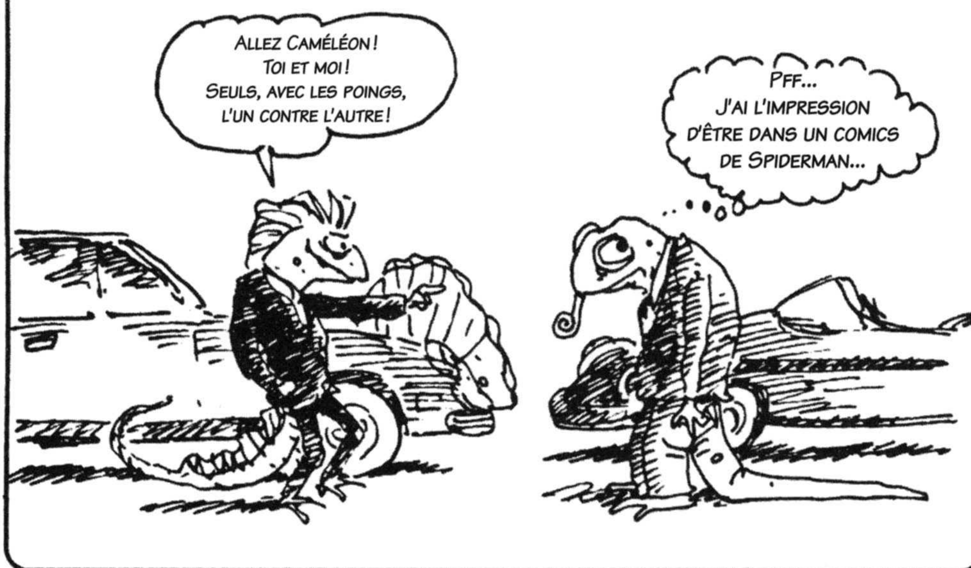
$$z_{OBS} = \frac{\bar{X}_1 - \bar{X}_2}{s(\bar{X}_1 - \bar{X}_2)}$$

ET LES VALEURS P FONCTIONNENT DE MANIÈRE HABITUELLE.



Et qu'en est-il de la comparaison des MOYENNES DE PETITS ÉCHANTILLONS ?

VOUS VOUS RAPPELEZ CAMÉLÉON AUTOMOBILES ? LEUR CONCURRENT, **IGUANE AUTOS,** AFFIRME QUE LEUR CAPOT AVEC DÉCORATION EN POLYSTYRÈNE OFFRE UNE MEILLEURE PROTECTION AUX CRASHS FRONTAUX ET ILS ONT CRASHÉ **SEPT IGUANE** POUR LE PROUVER !



LEURS RÉSULTATS COMPARÉS AUX CAMÉLÉON SONT :

CAMÉLÉON

1	150 €
2	400 €
3	720 €
4	500 €
5	930 €
n_1	5
$\bar{x}_1$	540 €
s_1	299 €

IGUANE

1	50 €
2	200 €
3	150 €
4	400 €
5	750 €
6	400 €
7	150 €
n_2	7
$\bar{x}_2$	300 €
s_2	238 €



LE t DE STUDENT PEUT ÊTRE UTILISÉ SI LES DEUX POPULATIONS ONT UNE FORME DE MONTICULE ET ONT LE MÊME ÉCART-TYPE $\sigma = \sigma_1 = \sigma_2$. LE SEUL HIC EST QUE NOUS DEVONS FAIRE UNE **MOYENNE PONDÉRÉE** DES VARIANCES D'ÉCHANTILLON POUR ESTIMER σ^2.

$$s_{EST}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$



L'ERREUR-TYPE EST LA MÊME QUE POUR DE GRANDS ÉCHANTILLONS, SAUF QUE s_{EST} REMPLACE s_1 ET s_2.

$$s(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_{EST}^2}{n_1} + \frac{s_{EST}^2}{n_2}}$$

$$= s_{EST}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

L'INTERVALLE AVEC $1 - \alpha$ DE CONFIANCE EST ALORS :

$$\mu_1 - \mu_2 \in \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\, s(\bar{X}_1 - \bar{X}_2)$$

OÙ $t_{\alpha/2}$ EST LE FRACTILE D'UN t DE STUDENT AVEC $n_1 + n_2 - 2$ DEGRÉS DE LIBERTÉ.

LES CONSTRUCTEURS REPTILIENS D'AUTOMOBILE CONVIENNENT QUE LEURS ÉCARTS-TYPES SONT PROCHES ET QUE LEURS HISTOGRAMMES DE RÉPARATIONS ONT UNE FORME DE CLOCHE. ILS CALCULENT :

$$s_{EST} = \sqrt{\frac{4 \times 299^2 + 6 \times 238^2}{10}} = 264$$

$$s(\bar{X}_1 - \bar{X}_2) = 264\sqrt{\frac{1}{5} + \frac{1}{7}} = 155$$

L'INTERVALLE DE CONFIANCE À 95 % EST :

$$\mu_1 - \mu_2 \in 540 - 300 \pm t_{0,025} \times 155$$

$$\mu_1 - \mu_2 \in 240 \pm 2,23 \times 155$$

$$\mu_1 - \mu_2 \in \mathbf{240 \pm 346}$$

COMME CELA INCLUT LA VALEUR 0, IGUANE AUTOS **N'A PAS** PROUVÉ LA PRÉSENCE D'AMÉLIORATIONS SIGNIFICATIVES POUR LES COÛTS DE RÉPARATION.



Le prochain exemple montre l'inconvénient de suivre bêtement le livre de recettes : le propriétaire d'une large flotte de taxis américains veut comparer les distances parcourues, en miles, avec, soit une **essence A**, soit une **essence B.**



Il sélectionne 100 taxis, et leur assigne au hasard l'une des essences. Après une journée type de taxi, il obtient :

	Taille d'échantillon	Moyenne de miles	Écart-type
A	50	25	5,00
B	50	26	4,00



La différence d'échantillon est :

$\bar{x}_1 - \bar{x}_2 = 25 - 26 = -1$

L'essence B est-elle meilleure que l'essence A ?



En raison des écarts-types élevés, l'erreur-type est assez substantielle.

$$s(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= \sqrt{\frac{25}{50} + \frac{16}{50}}$$

$$= \mathbf{0{,}905}$$

Avec 95 % de confiance, nous avons :

$$\mu_1 - \mu_2 \in \bar{x}_1 - \bar{x}_2 \pm z_{0{,}025} s(\bar{X}_1 - \bar{X}_2)$$

$$\mu_1 - \mu_2 \in -1 \pm (1{,}96 \times 0{,}905)$$

$$\mu_1 - \mu_2 \in \mathbf{-1 \pm 1{,}774}$$

Cela inclut la valeur 0, qui correspond à $\mu_1 = \mu_2$.

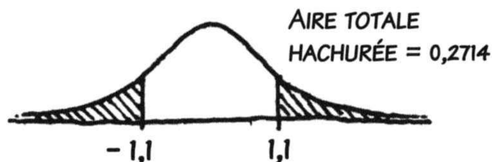


La valeur P pour l'hypothèse alternative, $H_a : \mu_1 \neq \mu_2$, vaut :

$$P(|z| > |z_{OBS}|) = P\left(|z| > \frac{1}{0{,}905}\right)$$

$$= P(|z| > 1{,}1) = 2 \times 0{,}1357$$

$$= \mathbf{0{,}2714}$$



Cela excède le seuil de signification $\alpha = 0{,}05$. Nous en concluons donc que les preuves en faveur de l'essence B sont très faibles.

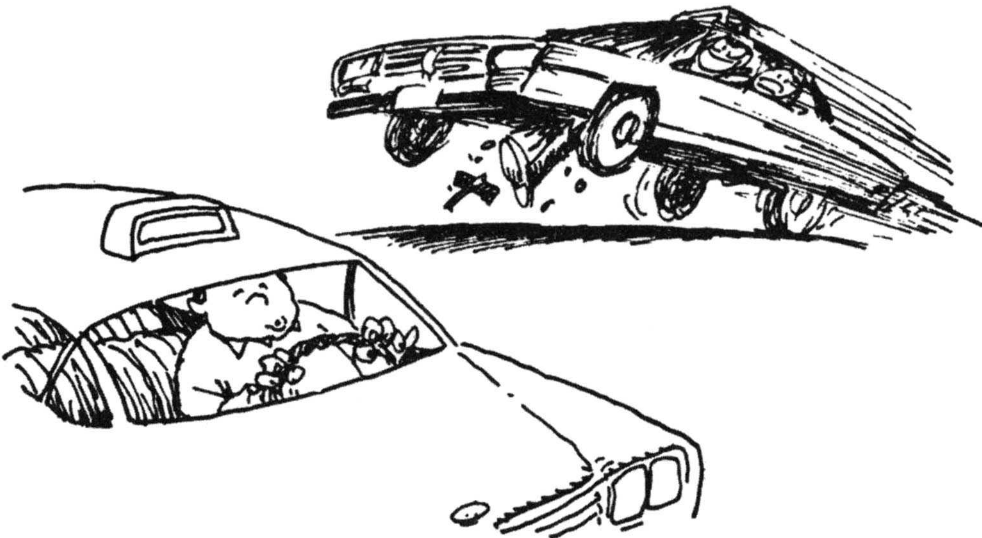


COMPARAISON APPARIÉE
Une meilleure façon pour comparer les essences



LE PROPRIÉTAIRE DE TAXIS A SUIVI LE LIVRE À LA LETTRE. SES ÉCHANTILLONS ÉTAIENT ALÉATOIRES D'UNE TAILLE SUFFISAMMENT GRANDE. IL A SIMPLEMENT OUBLIÉ DE **RÉFLÉCHIR** QUAND C'ÉTAIT NÉCESSAIRE.

BIEN QUE L'ESSENCE B SEMBLE LÉGÈREMENT SUPÉRIEURE À L'ESSENCE A, L'INTERVALLE DE CONFIANCE ÉTAIT LARGE À CAUSE DES GRANDS ÉCARTS-TYPES (AINSI, **LES KILOMÈTRES PARCOURUS VARIAIENT BEAUCOUP D'UN TAXI À L'AUTRE**). POURQUOI TANT DE VARIABILITÉ ? PARCE QUE LES TAXIS ET LEURS CONDUCTEURS ONT DES **CARACTÉRISTIQUES** DIFFÉRENTES.



Une bien meilleure façon d'aborder cette étude est d'affecter l'essence A et B au **même taxi** sur des **jours différents.**



Le traitement peut encore être aléatoire avec un lancer de pièce pour décider si on utilise l'essence A le mardi ou le mercredi. Et on peut même économiser l'argent et le temps du propriétaire en ne pratiquant l'expérience que sur 10 taxis !



Taxi	Essence A	Essence B	Différence
1	27,01	26,95	0,06
2	20,00	20,44	− 0,44
3	23,41	25,05	− 1,64
4	25,22	26,32	− 1,10
5	30,11	29,56	0,55
6	25,55	26,60	− 1,05
7	22,23	22,93	− 0,70
8	19,78	20,23	− 0,45
9	33,45	33,95	− 0,50
10	25,22	26,01	− 0,79
Moyenne	25,198	25,804	− 0,61
Écart-type	4,27	4,10	0,61

Notez que les moyennes et écarts-types des essences A et B sont à peu près les mêmes que précédemment. C'était prévisible puisque ces statistiques ont les mêmes sources de variation que des échantillons non appariés. Mais maintenant la **colonne des différences** a un faible écart-type. Ces différences éliminent la **variabilité entre les taxis** en comparant les performances de l'essence pour une même voiture.

LA DIFFÉRENCE d_i FOURNIT **UNE MESURE DE LA DIFFÉRENCE POUR CHAQUE TAXI** ET NOUS POUVONS L'UTILISER POUR NOTRE STATISTIQUE DE TEST t SUR DE PETITS ÉCHANTILLONS.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

OÙ $\bar{d}$ EST LA MOYENNE DES DIFFÉRENCES DE L'ÉCHANTILLON (– 0,61 DANS L'EXEMPLE) ET s_d EST LEUR ÉCART-TYPE (ICI 0,61).

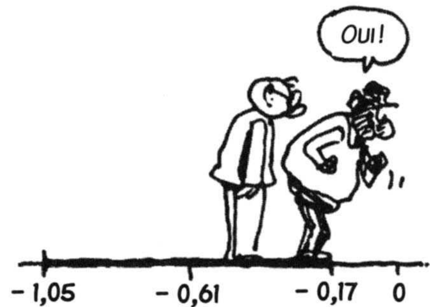


L'INTERVALLE DE CONFIANCE À 95 % CENTRÉ EN $\bar{d}$ EST :

$$\mu_d \in \bar{d} \pm t_{0,025} \times \frac{s_d}{\sqrt{n}}$$

$$\mu_d \in -0,61 \pm 2,26 \times \frac{0,61}{\sqrt{10}}$$

$$\mu_d \in \mathbf{-0,61 \pm 0,44}$$



NOUS SOMMES SÛRS À 95 % QUE $-1,05 \leq \mu_d \leq -0,17$, CE QUI CONSTITUE UNE PREUVE SOLIDE QUE L'ESSENCE B EST SUPÉRIEURE.

LA VALEUR P DU TEST D'HYPOTHÈSE PEUT ÊTRE CALCULÉE PAR ORDINATEUR.

$$H_a : \mu_d \neq 0$$

$$\text{VALEUR DE } P = P(|t| \geq |t_{OBS}|)$$

$$= P\left(|t| \geq \frac{0,61}{0,19}\right)$$

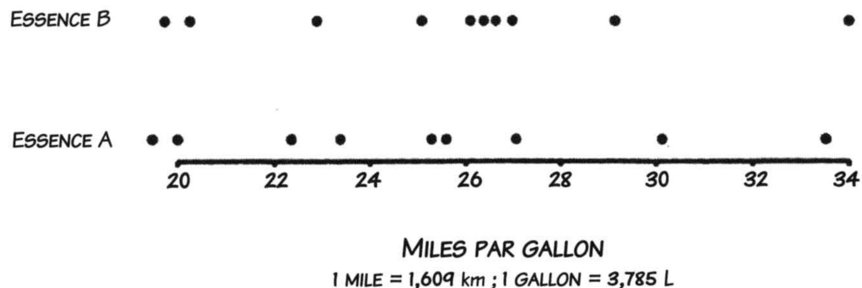
$$= P(|t| \geq 3,21)$$

$$\mathbf{= 0,011}$$



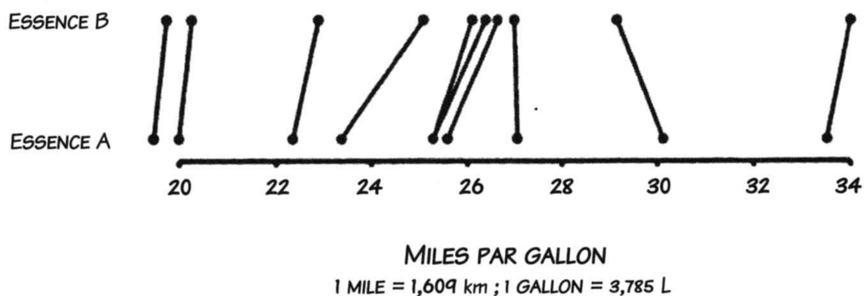
LA VALEUR P EST INFÉRIEURE À 0,05, DONC, À NOUVEAU, L'ESSENCE B PASSE LE TEST.

VOICI LE DIAGRAMME EN POINTS DES DONNÉES EN MILES PARCOURUS SELON L'ESSENCE : LE PREMIER DESSIN MONTRE LES DONNÉES NON APPARIÉES.



MILES PAR GALLON
1 MILE = 1,609 km ; 1 GALLON = 3,785 L

ET VOICI LES MÊMES DONNÉES APPARIÉES PAR TAXI.

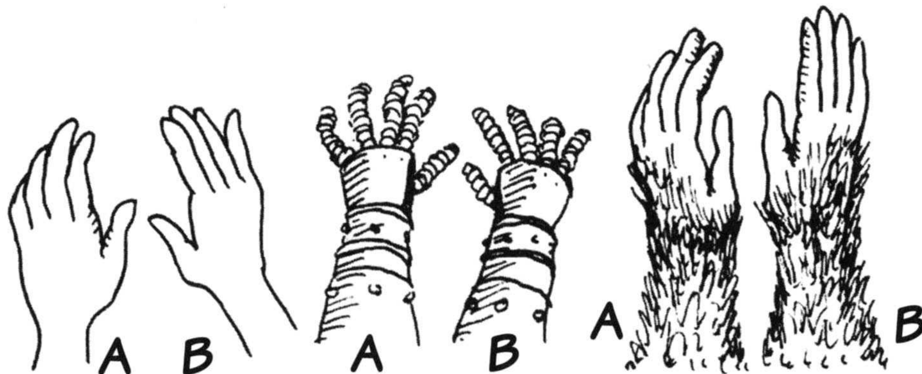


MILES PAR GALLON
1 MILE = 1,609 km ; 1 GALLON = 3,785 L

LA PRÉDOMINANCE DE SEGMENTS INCLINÉS VERS LA DROITE EST UN INDICE FORT QUE L'ESSENCE B ACCROÎT LA DISTANCE PARCOURUE.



UTILISER DES ÉCHANTILLONS APPARIÉS EST LA MÉTHODE LA PLUS EFFICACE POUR RÉDUIRE LA VARIABILITÉ NATURELLE QUAND ON COMPARE DES TRAITEMENTS. PAR EXEMPLE, SI L'ON COMPARE DEUX MARQUES DE CRÈMES POUR LES MAINS, ON PEUT ALÉATOIREMENT AFFECTER LA MAIN DROITE À L'UNE ET LA MAIN GAUCHE À L'AUTRE POUR LE MÊME SUJET. ON ÉLIMINE AINSI LA VARIABILITÉ DUE À DES TYPES DE PEAUX DIFFÉRENTS.

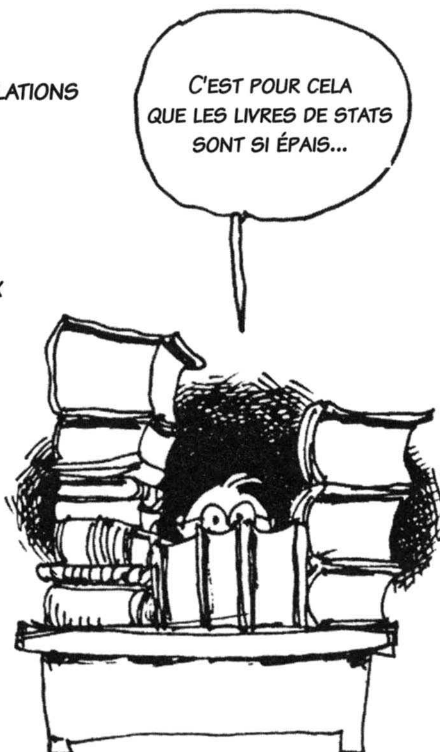


OU SI ON COMPARE DEUX MARQUES DE CÉRÉALES, CHAQUE « GOÛTEUR » ÉVALUERA LES DEUX CÉRÉALES (DANS UN ORDRE ALÉATOIRE). UNE COMPARAISON APPARIÉE SUPPRIME LE BIAIS NATUREL D'UN GOÛTEUR QUI SERAIT POUR OU CONTRE LES CÉRÉALES D'UNE MANIÈRE GÉNÉRALE.



DANS CE CHAPITRE, NOUS AVONS APPLIQUÉ LES IDÉES DE BASE SUR LES INTERVALLES DE CONFIANCE ET LES TESTS D'HYPOTHÈSES À LA COMPARAISON DE DEUX POPULATIONS. IL EXISTE D'INNOMBRABLES AUTRES POSSIBILITÉS. NOUS AURIONS PU POURSUIVRE EN DÉCRIVANT DES COMPARAISONS :

- D'ÉCARTS-TYPES DE DEUX POPULATIONS AVEC DE PETITS ÉCHANTILLONS ;
- DE MOYENNES DE PLUS DE DEUX POPULATIONS AVEC DE GRANDS ÉCHANTILLONS ;
- DE MOYENNES DE PLUS DE DEUX POPULATIONS AVEC DE PETITS ÉCHANTILLONS ;



ETC. !

DANS LA PRATIQUE, LES STATISTICIENS DÉTERMINENT LA NATURE GÉNÉRALE D'UN PROBLÈME ET CONSULTENT ALORS LE LIVRE DE RÉFÉRENCE ADAPTÉ.



LA SEULE IDÉE VÉRITABLEMENT NOUVELLE DE CE CHAPITRE ÉTAIT L'**APPARIEMENT** POUR RÉALISER DES TESTS DE **COMPARAISON.** DANS LE PROCHAIN CHAPITRE, NOUS VERRONS D'AUTRES TYPES DE DISPOSITIFS EXPÉRIMENTAUX.


VOUS VOULEZ ACHETER UN CAMÉLÉON D'OCCASION ?

Chapitre 10
Méthodes expérimentales

LA CONCEPTION EST SOUVENT CE QUI GARANTIT LE SUCCÈS OU L'ÉCHEC D'UNE EXPÉRIENCE. DANS L'EXEMPLE DE LA COMPARAISON APPARIÉE, NOTRE STATISTICIEN A INVERSÉ LES RÔLES PASSANT D'UNE COLLECTE ET D'UNE ANALYSE PASSIVE DES DONNÉES À UNE PARTICIPATION ACTIVE DANS LE DISPOSITIF DE L'EXPÉRIENCE.

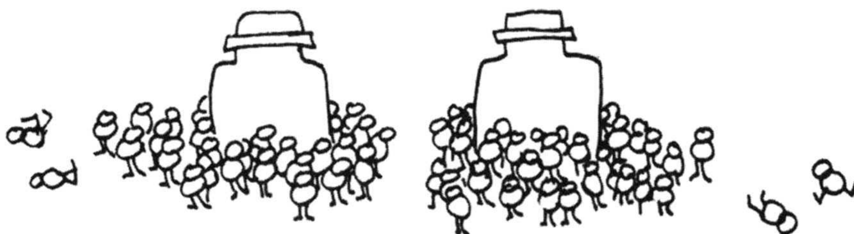


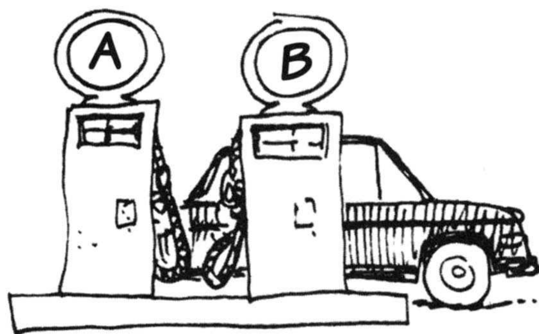
DANS CE CHAPITRE, NOUS INTRODUISONS LES IDÉES DE BASE DES MÉTHODES EXPÉRIMENTALES, EN LAISSANT LES ANALYSES NUMÉRIQUES DÉTAILLÉES À VOTRE LOGICIEL DE STATISTIQUES.



PAS DE FORMULES DANS CE CHAPITRE... DÉSOLÉ !

LES ÉLÉMENTS D'UN DISPOSITIF SONT : LES **UNITÉS EXPÉRIMENTALES** ET LES TRAITEMENTS QUI SERONT ASSIGNÉS AUX UNITÉS. L'OBJECTIF DU DISPOSITIF EST DE COMPARER LES TRAITEMENTS.





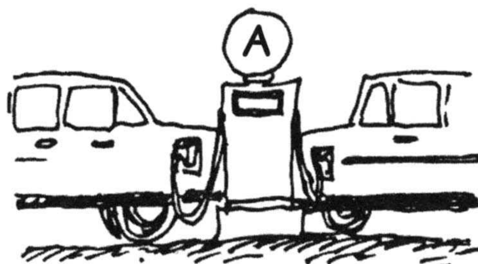
POUR DES ESSAIS MÉDICAUX, LES **PATIENTS** SONT LES UNITÉS, ET LES **MÉDICAMENTS** SONT LES TRAITEMENTS. DANS L'EXEMPLE SUR L'ESSENCE, LES UNITÉS EXPÉRIMENTALES SONT LES TAXIS, ET LES TRAITEMENTS À COMPARER SONT LES ESSENCES A ET B.

DANS DES EXPÉRIENCES AGRICOLES, LES UNITÉS EXPÉRIMENTALES SONT SOUVENT DES PARCELLES DE TERRAIN, ET LES TRAITEMENTS PEUVENT CORRESPONDRE À L'APPLICATION DE DIFFÉRENTES VARIÉTÉS DE BLÉ, DE PESTICIDES, DE FERTILISANTS, ETC.

AUJOURD'HUI, LES IDÉES DES MÉTHODES EXPÉRIMENTALES SONT LARGEMENT UTILISÉES DANS L'**OPTIMISATION DES PROCESSUS INDUSTRIELS,** EN **MÉDECINE** ET DANS LES **SCIENCES SOCIALES.** TOUT DISPOSITIF EXPÉRIMENTAL UTILISE **TROIS PRINCIPES DE BASE** QUI ÉTAIENT TOUS PRÉSENTS DANS L'EXEMPLE DES TAXIS.



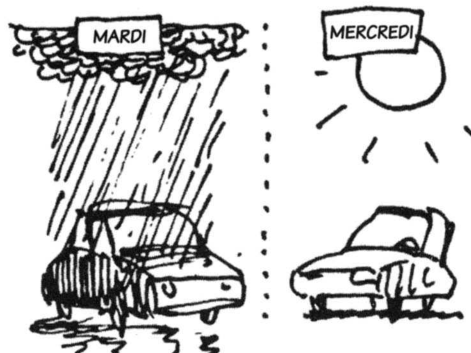
La réplication : LES MÊMES TRAITEMENTS SONT ASSIGNÉS À DIFFÉRENTES UNITÉS EXPÉRIMENTALES. SANS RÉPLICATION, IL EST IMPOSSIBLE D'ÉVALUER LA VARIABILITÉ NATURELLE ET LES ERREURS DE MESURE.



Le contrôle local : IL SE RÉFÈRE À TOUTE MÉTHODE QUI TIENT COMPTE DE LA VARIABILITÉ NATURELLE ET LA RÉDUIT. UN MOYEN EST DE REGROUPER LES UNITÉS EXPÉRIMENTALES SIMILAIRES EN BLOCS. DANS L'EXEMPLE DES TAXIS, CHACUN D'EUX UTILISAIT LES DEUX ESSENCES, ON DIT ALORS QUE LE TAXI EST UN BLOC.



La randomisation : C'EST L'ÉTAPE ESSENTIELLE POUR TOUTE ÉTUDE STATISTIQUE! LES TRAITEMENTS DOIVENT ÊTRE ASSIGNÉS ALÉATOIREMENT AUX UNITÉS EXPÉRIMENTALES. POUR CHAQUE TAXI, L'ESSENCE A EST UTILISÉE LE MARDI OU LE MERCREDI À PILE OU FACE. SI CE N'AVAIT PAS ÉTÉ LE CAS, LES RÉSULTATS AURAIENT PU ÊTRE FAUSSÉS PAR DES DIFFÉRENCES ENTRE LE MARDI ET LE MERCREDI.



SUPPOSONS MAINTENANT QUE L'ON VEUILLE ÉTUDIER LES EFFETS DE DEUX MARQUES DE PNEUS ET AUSSI DE DEUX ESSENCES. IL Y A QUATRE TRAITEMENTS POSSIBLES QUE NOUS POUVONS PRÉSENTER DANS UNE TABLE FACTORIELLE 2×2.
LES DEUX FACTEURS ÉTANT L'ESSENCE ET LE TYPE DE PNEU.

	ESSENCE A	ESSENCE B
PNEU A	*a*	*b*
PNEU B	*c*	*d*



ON PEUT ASSIGNER LES QUATRE TRAITEMENTS DE FAÇON ALÉATOIRE SUR QUATRE JOURS DIFFÉRENTS POUR CHAQUE TAXI. LES QUATRE TRAITEMENTS (*a, b, c et d*) SONT RÉPÉTÉS DANS CHAQUE BLOC (TAXI). IL S'AGIT ALORS D'UN **DISPOSITIF EN BLOCS ALÉATOIRES COMPLETS.**

JUSQU'À PRÉSENT, NOUS AVONS FAIT L'HYPOTHÈSE QUE CHAQUE JOUR DE LA SEMAINE EST IDENTIQUE. MAIS ON PEUT CONTRÔLER CELA AUSSI EN UTILISANT SEULEMENT QUATRE TAXIS ET EN ASSIGNANT UN TRAITEMENT SELON LE PLAN DU TABLEAU DE DROITE.

		JOUR 1	2	3	4
TAXI	1	*a*	*b*	*c*	*d*
	2	*b*	*c*	*d*	*a*
	3	*c*	*d*	*a*	*b*
	4	*d*	*a*	*b*	*c*



UN TABLEAU 4 × 4 AVEC 4 ÉLÉMENTS DIFFÉRENTS, CHACUN APPARAISSANT UNE SEULE FOIS DANS CHAQUE LIGNE ET COLONNE, S'APPELLE UN **carré latin.** DANS CETTE EXPÉRIENCE, LES QUATRE JOURS ET LES QUATRE TAXIS REÇOIVENT LES QUATRE TRAITEMENTS EXACTEMENT UNE FOIS CHACUN.

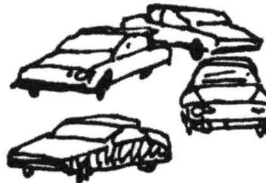




L'ÉTAPE DE RANDOMISATION SÉLECTIONNE UN CARRÉ LATIN AU HASARD DE LA LISTE DE TOUS LES CARRÉS LATINS POSSIBLES À QUATRE ÉLÉMENTS.

SI QUATRE UNITÉS NE SONT PAS SUFFISANTES, ON PEUT AUGMENTER LE NOMBRE D'UNITÉS EXPÉRIMENTALES EN **RÉPÉTANT** LE DISPOSITIF EXPÉRIMENTAL. SI ON PART DE HUIT TAXIS, ON PEUT LES DIVISER EN DEUX GROUPES DE QUATRE ET RÉPÉTER LE DISPOSITIF DANS CHAQUE GROUPE.







NOUS AVIONS PROMIS DE NE PAS ENTRER DANS LE DÉTAIL DE L'ANALYSE DE DONNÉES. MAIS VOICI EN GROS COMMENT GÉRER UN DISPOSITIF COMPLEXE DE CE TYPE.



L'ANALYSE DES DISPOSITIFS EXPÉRIMENTAUX SE FAIT EN ALLOUANT LA VARIABILITÉ TOTALE AUX DIFFÉRENTES SOURCES. DANS L'EXEMPLE DES TAXIS, LES SOURCES DE VARIATION SONT LE TAXI, LA MARQUE DE PNEU, LE TYPE D'ESSENCE, LE JOUR ET UNE ERREUR ALÉATOIRE. L'ANALYSE DE VARIANCE, OU **ANOVA** EN ABRÉGÉ, DIVISE LA VARIATION TOTALE EN PORTIONS POUR CHAQUE SOURCE.

DANS LE PROCHAIN CHAPITRE, NOUS EXPLIQUERONS EN DÉTAIL UN MODÈLE POUR ANALYSER DES DISPOSITIFS COMPLEXES : LE **MODÈLE DE RÉGRESSION LINÉAIRE.** AVEC LA RÉGRESSION LINÉAIRE, VOUS POURREZ VOIR ANOVA NUMÉRIQUEMENT ET DE PRÈS...



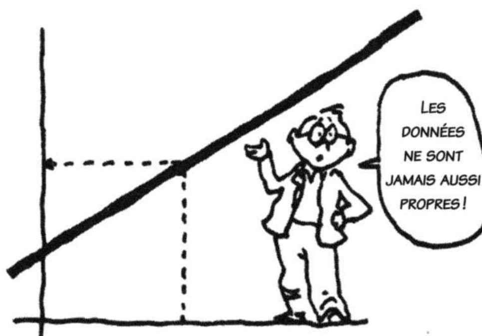
Chapitre 11
Régression linéaire

JUSQU'À PRÉSENT, NOUS AVONS FAIT DES STATISTIQUES SUR **UNE SEULE VARIABLE À LA FOIS,** QU'ELLE VIENNE D'UNE POPULATION DE PRENEURS DE PILULES OU DE VOITURES ACCIDENTÉES. DANS CE CHAPITRE, NOUS ALLONS VOIR COMMENT RELIER **DEUX** VARIABLES. CONNAISSANT LES POIDS DES 92 ÉTUDIANTS DU CHAPITRE 2, NOUS NOUS DEMANDONS COMMENT ILS SONT RELIÉS À LA TAILLE DE CES ÉTUDIANTS.

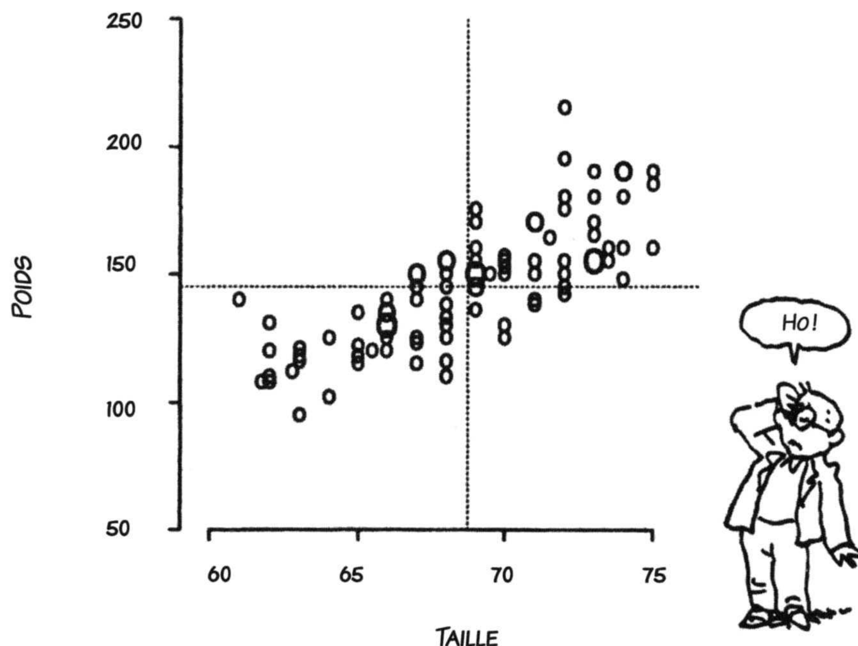


C'EST UN EXEMPLE D'UNE LARGE CLASSE DE QUESTIONS IMPORTANTES : LA **PRESSION ARTÉRIELLE** PRÉDIT-ELLE LA **DURÉE DE VIE ?** LES **RÉSULTATS DU BAC** PRÉDISENT-ILS LES **PERFORMANCES EN LICENCE ?** LIRE UN LIVRE DE STATISTIQUES FAIT-IL DE VOUS UNE **MEILLEURE PERSONNE ?**

EN CLASSE DE MATHÉMATIQUES, VOUS AVEZ PROBABLEMENT VU DES RELATIONS DÉFINIES COMME DES GRAPHIQUES. ÉTANT DONNÉ x, ON PEUT ALORS PRÉDIRE y. MAIS EN STATISTIQUES, LES CHOSES NE SONT PAS AUSSI SIMPLES ! ON SAIT (OU ON SUPPOSE) QUE LA TAILLE A UNE INCIDENCE SUR LE POIDS – MAIS CE N'EST PAS LA SEULE. IL Y A AUSSI D'AUTRES FACTEURS COMME LE SEXE, L'ÂGE, LE TYPE PHYSIQUE, ET UNE **COMPOSANTE ALÉATOIRE.**



POUR CE CHAPITRE, NOTONS y LA DONNÉE DE POIDS EN LIVRES ET x LA DONNÉE DE TAILLE EN POUCES. AINSI (x_i, y_i) EST LA TAILLE ET LE POIDS DE L'ÉTUDIANT i. ON PEUT TRACER LES POINTS (x_i, y_i) DANS UN PLAN, ON APPELLE CELA **UN NUAGE DE POINTS.**



(CERTAINS POINTS SONT PLUS GROS CAR ILS REPRÉSENTENT DEUX OU TROIS ÉTUDIANTS AYANT LE MÊME POIDS ET LA MÊME TAILLE.)

POUVONS-NOUS PRÉVOIR LE POIDS y D'UN ÉTUDIANT CONNAISSANT SA TAILLE x ?

L'ANALYSE DE RÉGRESSION

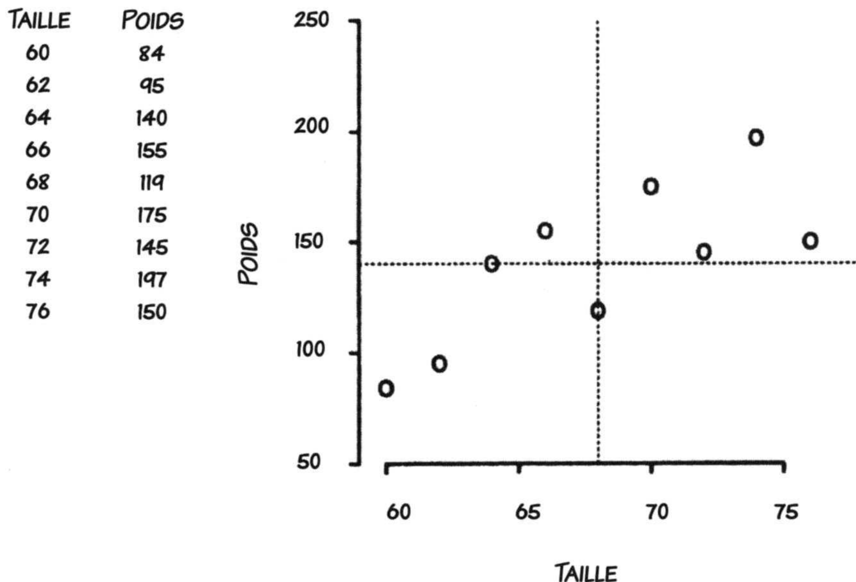
L'ANALYSE DE RÉGRESSION CONSISTE À AJUSTER UNE LIGNE DROITE À CE NUAGE DÉSORDONNÉ DE POINTS. x EST APPELÉE VARIABLE **INDÉPENDANTE** OU **EXPLICATIVE**, ET y EST LA VARIABLE **DÉPENDANTE** OU DE **RÉPONSE**. LA DROITE DE **RÉGRESSION AFFINE** A LA FORME SUIVANTE :

$$y = ax + b$$



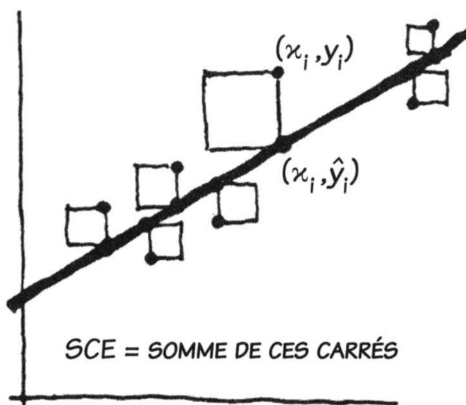
POUR ILLUSTRER LA MÉTHODE D'AJUSTEMENT, UTILISONS UN SOUS-ENSEMBLE UN PEU ARRANGÉ DE DONNÉES CONTENANT SEULEMENT NEUF OBSERVATIONS D'ÉTUDIANTS.

TAILLE	POIDS
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



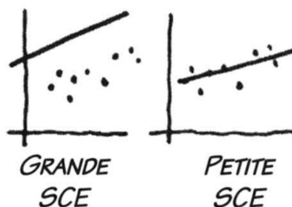
COMMENT OBTENIR LA DROITE LA MIEUX AJUSTÉE AUX DONNÉES ?

L'IDÉE EST DE **MINIMISER** L'ÉCART TOTAL DES VALEURS y OBSERVÉES PAR RAPPORT À LA DROITE. COMME AVEC LA VARIANCE, ON REGARDE PLUTÔT LA **DISTANCE AU CARRÉ** ENTRE y ET LA DROITE, ON LES SOMME POUR OBTENIR **LA SOMME DES CARRÉS DES ERREURS** OU **RÉSIDUS**.



$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

C'EST UNE MESURE AGRÉGÉE QUI CALCULE DE COMBIEN LA DROITE « PRÉDISANT y_i » (C'EST-À-DIRE $\hat{y}_i$) DIFFÈRE DES DONNÉES (C'EST-À-DIRE DES VALEURS OBSERVÉES y_i).



LA RÉGRESSION OU DROITE DES MOINDRES CARRÉS

EST LA DROITE AVEC LA PLUS **PETITE** SCE.



NOTE HISTORIQUE : POURQUOI APPELLE-T-ON CETTE PROCÉDURE L'**ANALYSE DE RÉGRESSION** ? DANS LES ANNÉES 1880, LE GÉNÉTICIEN FRANCIS GALTON (1822-1911) DÉCOUVRIT LE PHÉNOMÈNE DE **RÉGRESSION VERS LA MOYENNE.** EN RECHERCHANT LES RÈGLES DE L'HÉRÉDITÉ, IL CONSTATA QUE LA TAILLE DES FILS TENDAIT À **RÉGRESSER** VERS LA TAILLE MOYENNE DE LA POPULATION, PAR RAPPORT À LA TAILLE DES PÈRES. LES PÈRES DE GRANDE TAILLE TENDAIENT À AVOIR DES FILS PLUS PETITS ET VICE VERSA. GALTON DÉVELOPPA L'ANALYSE DE RÉGRESSION POUR ÉTUDIER CE QU'IL APPELA UNE « RÉGRESSION VERS LA MÉDIOCRITÉ ».



SANS TOURNER AUTOUR DU POT, NOUS DONNONS LA FORMULE DE LA DROITE DE RÉGRESSION SANS LA DÉMONTRER. C'EST COMPLIQUÉ MAIS CALCULABLE.

$$y = ax + b$$

OÙ

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

ET

$$a = \bar{y} - b\bar{x}$$

OÙ $\bar{x}$ ET $\bar{y}$ SONT LES MOYENNES DE $\{x_i\}$ ET $\{y_i\}$ RESPECTIVEMENT. LES SOMMES SONT POUR i ALLANT DE 1 À n.



PARCE QUE NOUS ALLONS RETROUVER CERTAINES EXPRESSIONS, ON VA LES ABRÉGER (DE NOUVEAU, LA SOMME SE FAIT POUR TOUT i, SAUF LORSQUE C'EST PRÉCISÉ).

$$SC_{xx} = \sum (x_i - \bar{x})^2$$

$$SCT = SC_{yy} = \sum (y_i - \bar{y})^2$$

SOMME DES CARRÉS DES ÉCARTS À LA MOYENNE POUR x_i ET y_i. *SCT* VEUT DIRE SOMME DES CARRÉS TOTAUX.

$$SC_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

LE PRODUIT CROISÉ ET SC_{xx} DÉTERMINENT LE COEFFICIENT b.



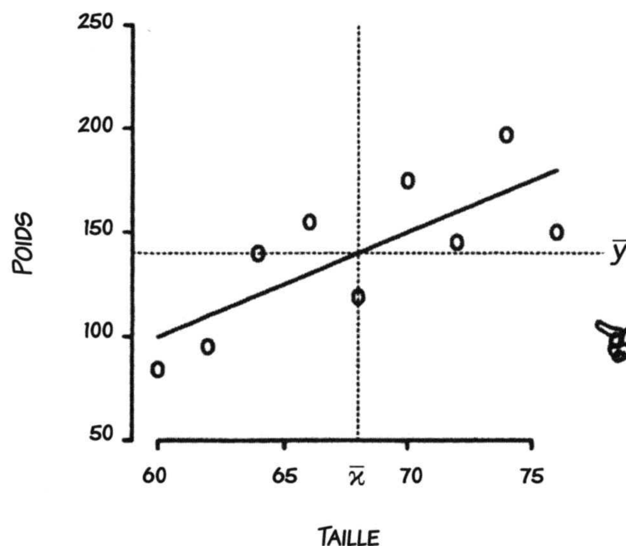
POUR LES DONNÉES ARRANGÉES, VOICI LES CALCULS PAS À PAS :

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
60	84	-8	-56	64	3136	448
62	95	-6	-45	36	2025	270
64	140	-4	0	16	0	0
66	155	-2	15	4	225	-30
68	119	0	-21	0	441	0
70	175	2	35	4	1225	70
72	145	4	5	16	25	20
74	197	6	57	36	3249	342
76	150	8	10	64	100	80
SOMME 612	1260			$SC_{xx} = 240$	$SC_{yy} = 10\,426$	$SC_{xy} = 1200$
$\bar{x} = 68$	$\bar{y} = 140$					

LA DROITE DE RÉGRESSION SE DÉTERMINE À PARTIR DES RÉSUMÉS STATISTIQUES EN BAS DU TABLEAU.

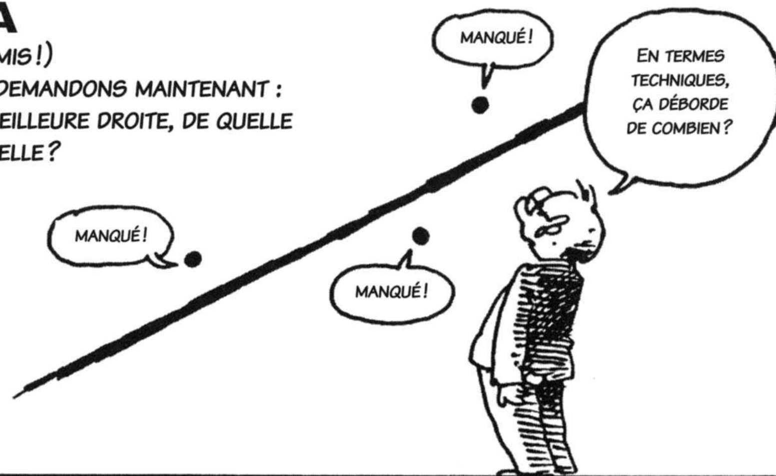
$$b = \frac{SC_{xy}}{SC_{xx}} = \frac{1200}{240} = \mathbf{5} \qquad a = \bar{y} - b\bar{x} = 140 - 5 \times 68 = \mathbf{-200}$$

DONC $y = \mathbf{-200 + 5}$

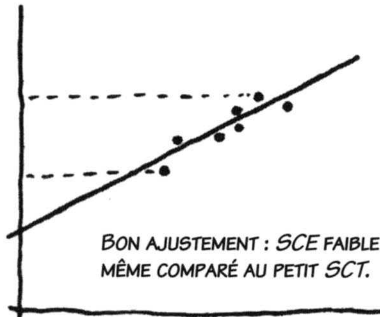


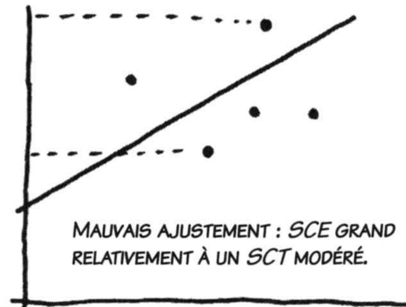
ANOVA

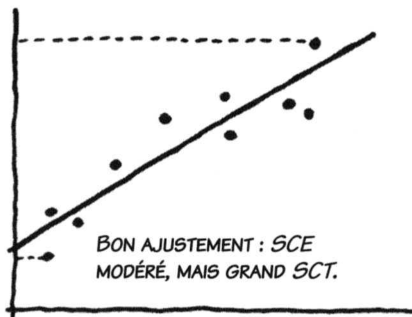
(COMME PROMIS !)
NOUS NOUS DEMANDONS MAINTENANT : SI C'EST LA MEILLEURE DROITE, DE QUELLE QUALITÉ EST-ELLE ?

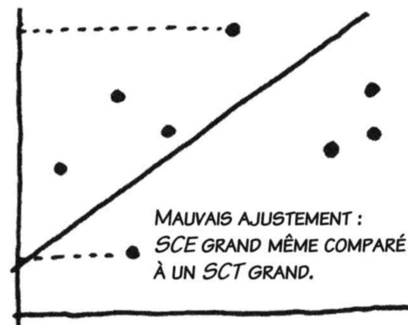


COMME VOUS POUVEZ L'IMAGINER, LA RÉPONSE À CETTE QUESTION DÉPEND DE LA FAÇON DONT LES DONNÉES SONT ÉPARPILLÉES : DE L'IMPORTANCE DE **SCE,** LA SOMME DES CARRÉS, RELATIVEMENT AUX **ÉCARTS TOTAUX** DES DONNÉES. QUELQUES EXEMPLES :









QUANTIFIONS CELA EN **SCINDANT LA VARIATION** EN y. REGARDEZ LA FIGURE DE DROITE POUR SUIVRE. NOUS AVONS

$$\hat{y}_i = a + bx_i$$

OÙ $\hat{y}_i$ EST LE POIDS **ESTIMÉ** PAR LA DROITE DE RÉGRESSION.

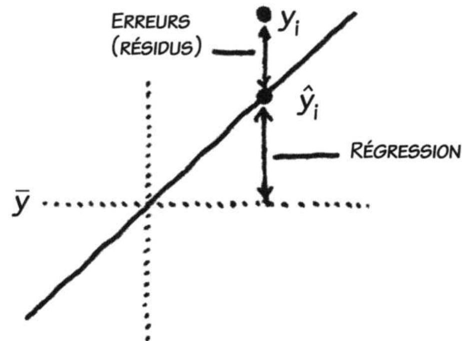


Table ANOVA

SOURCE DE VARIATION	SOMMES DES CARRÉS	VALEURS POUR LES DONNÉES
RÉGRESSION	$SCR = \sum (\hat{y}_i - \bar{y})^2$	6 000
ERREURS (RÉSIDUS)	$SCE = \sum (y_i - \hat{y}_i)^2$	4 426
TOTAL	$SCT = \sum (y_i - \bar{y})^2$	10 426

(À PROPOS, CE N'EST PAS ÉVIDENT QUE $SCT = SCR + SCE$, MAIS C'EST VRAI !) SINON, VOICI LE CALCUL DÉTAILLÉ DES SOMMES DE CARRÉS POUR LA RÉGRESSION ET LES RÉSIDUS POUR NOS DONNÉES AVEC $y = -200 + 5x$.

			RÉGRESSION		ERREURS (RÉSIDUS)	
x_i	y_i	$\hat{y}_i$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
60	84	100	−40	1 600	−16	256
62	95	110	−30	900	−15	225
64	140	120	−20	400	20	400
66	155	130	−10	100	25	625
68	119	140	0	0	−21	441
70	175	150	10	100	25	625
72	145	160	20	400	−15	225
74	197	170	30	900	27	729
76	150	180	40	1 600	−30	900
$\bar{x} = 68$	$\bar{y} = 140$		$SCR = 6\,000$		$SCE = 4\,426$	

SCR MESURE LA VARIATION TOTALE DUE À LA RÉGRESSION (LES VALEURS PRÉDITES DE y). SCE NOUS L'AVONS DÉJÀ VU. NOTEZ QUE

$$\frac{SCE}{SCT}$$

EST LA PROPORTION DES ERREURS (OU RÉSIDUS) RELATIVEMENT AUX ÉCARTS TOTAUX.



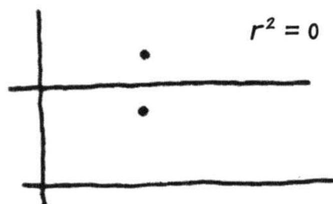
LE COEFFICIENT DE DÉTERMINATION

EST LA PROPORTION DE LA SCT IMPUTABLE À LA RÉGRESSION :

$$r^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$



(CAR SCR = SCT - SCE). r^2 EST TOUJOURS INFÉRIEUR OU ÉGAL À 1. PLUS C'EST PROCHE DE 1, PLUS LES OBSERVATIONS SONT PROCHES DE LA DROITE. $r^2 = 1$ INDIQUE QUE LES POINTS SONT ALIGNÉS SUR LA DROITE.



SI ON CALCULE, POUR NOTRE PETIT JEU DE DONNÉES, ON OBTIENT :

$$r^2 = \frac{6\,000}{10\,426} = \mathbf{0{,}58}$$

58 % DE LA VARIATION DU POIDS S'EXPLIQUE PAR LA TAILLE. LES 42 % RESTANTS SONT UN TERME D'« ERREUR ».



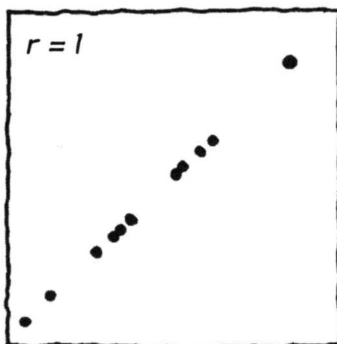
DE FAÇON ALTERNATIVE,
LE **coefficient de corrélation**
EST DÉFINI COMME LA RACINE
DE r^2 MULTIPLIÉE PAR LE SIGNE DE b.

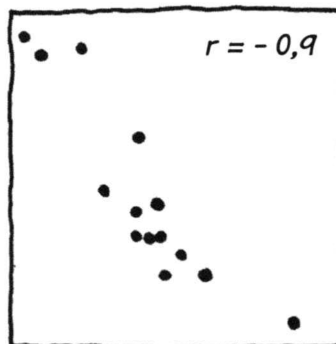
$$r = (\text{signe de } b)\sqrt{r^2}$$

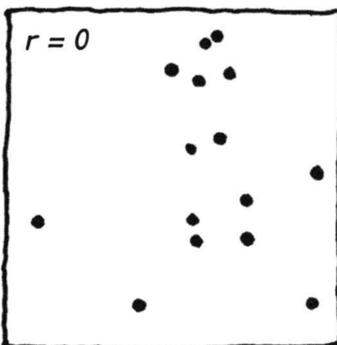
AINSI r EST POSITIF SI LA DROITE
EST CROISSANTE ET NÉGATIF
SI CELLE-CI EST DÉCROISSANTE.

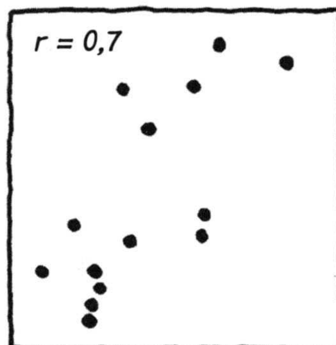


r MESURE À LA FOIS LA JUSTESSE DE L'AJUSTEMENT ET LE SENS DE LA RELATION ENTRE x ET y (EN INDIQUANT SI UN ACCROISSEMENT DE x AUGMENTE OU DIMINUE y).









MAINTENANT SOYONS HONNÊTES : PERSONNE OU PRESQUE NE FAIT CES CALCULS À LA MAIN DÉSORMAIS. AVEC UN ORDINATEUR, TOUT LE TRAVAIL PEUT S'EFFECTUER AVEC UNE LIGNE DE CODE.



SI ON UTILISE LE LOGICIEL MINITAB, DÉVELOPPÉ À PENN STATE, LA LIGNE DE COMMANDE EST :

MTB > régression 'poids' avec 1 variable indépendante 'taille'

ET LES RÉSULTATS SONT :



L'équation de la régression est

Poids = −200 + 5,00 * taille

Prédicteur	Coefficient	Écart-type	t Student	Valeur p
Constante	−200,0	110,7	−1,81	0,114
Taille	5,000	1,623	3,08	0,018

s = 25,15   R-carré = 57,5 %   R-carré (ajusté) = 51,5 %

Analyse de Variance

SOURCE	DL	SC	MC	F	Valeur p
Régression	1	6 000,0	6 000,0	9,49	0,018
Erreur	7	4 426,0	632,3		
Total	8	10 426,0			



MAINTENANT, FAISONS LA MÊME CHOSE AVEC LES VRAIES DONNÉES DES 92 ÉTUDIANTS :

MTB > régression 'poids' avec 1 variable indépendante 'taille'

ET LES RÉSULTATS SONT :

L'équation de la régression est

Poids = −205 + 5,09 * taille

Prédicteur	Coefficient	Écart-type	t Student	Valeur p
Constante	−204,74	29,16	−7,02	0,000
Taille	5,0918	0,4237	12,02	0,000

s = 14,79   R-carré = 61,6 %   R-carré (ajusté) = 61,2 %

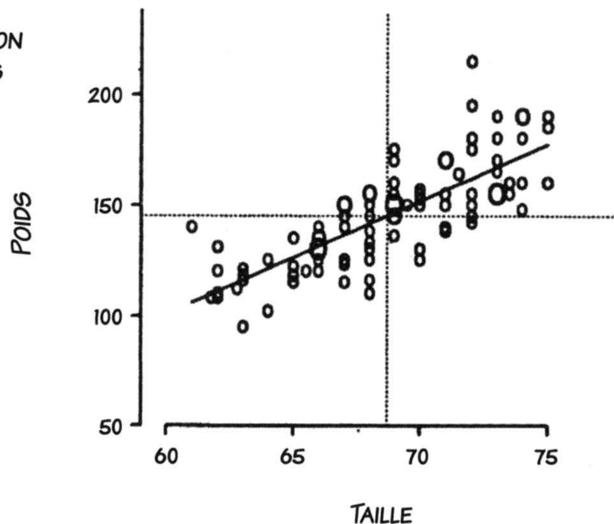
Analyse de Variance

SOURCE	DL	SC	MC	F	Valeur p
Régression	1	31592	31592	144,38	0,000
Erreur	7	19692	219		
Total	8	51284			

VOICI LE NUAGE DE POINTS AVEC LA DROITE ESTIMÉE. LE COEFFICIENT DE CORRÉLATION POUR NOTRE JEU DE DONNÉES VAUT

$r = +\sqrt{0,616} = \mathbf{0,78}$





INFÉRENCE STATISTIQUE

JUSQU'À MAINTENANT, NOUS AVONS FAIT DE L'**ANALYSE DE DONNÉES**, EN DÉTERMINANT LA MEILLEURE RELATION LINÉAIRE ENTRE LES DONNÉES OBSERVÉES DE x ET y. CHANGEONS MAINTENANT DE POINT DE VUE ET CONSIDÉRONS LES 92 ÉTUDIANTS COMME UN ÉCHANTILLON DE LA POPULATION GLOBALE DES ÉTUDIANTS. QUE POUVONS-NOUS EN DÉDUIRE ?



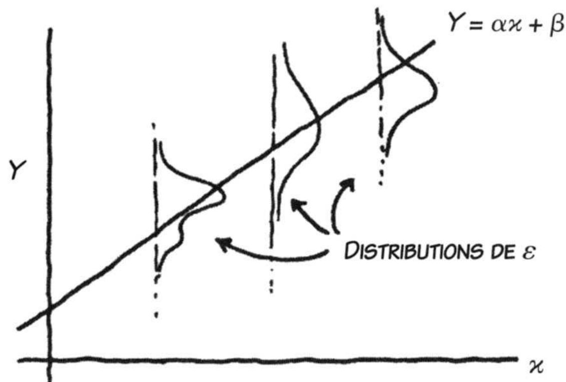
UN MODÈLE DE RÉGRESSION POUR LA POPULATION GLOBALE EST UNE RELATION LINÉAIRE

$$Y = \alpha + \beta x + \varepsilon$$

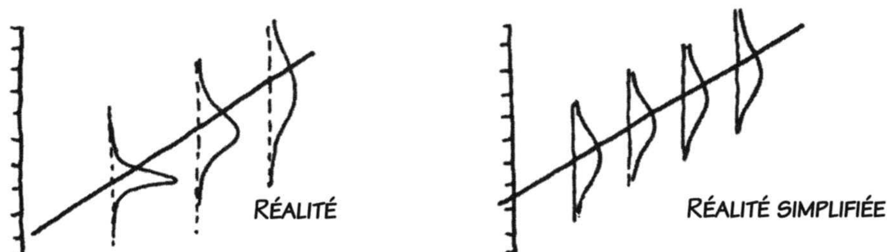
NOTEZ QUE LES LETTRES GRECQUES REPRÉSENTENT LES PARAMÈTRES DU MODÈLE !

Y EST LA VARIABLE ALÉATOIRE DÉPENDANTE ; x EST LA VARIABLE INDÉPENDANTE (QUI PEUT ÊTRE OU NON ALÉATOIRE) ; α ET β SONT LES PARAMÈTRES INCONNUS À ESTIMER ; ET ε REPRÉSENTE LES ERREURS ALÉATOIRES FLUCTUANTES.

POUR LE MODÈLE DE POIDS ET DE TAILLE, Y EST LE POIDS, x EST LA TAILLE, α ET β SONT INCONNUES. ON PEUT VOIR ε COMME UNE **COMPOSANTE ALÉATOIRE** DES POIDS Y POUR CHAQUE VALEUR DE TAILLE x.



LA DISTRIBUTION DE ε N'EST EN FAIT PAS LA MÊME POUR DES VALEURS DIFFÉRENTES DE x. LES PERSONNES DE 1,50 m VARIENT MOINS EN POIDS QUE CELLES DE 1,80 m. NÉANMOINS, NOUS FAISONS MAINTENANT UNE HYPOTHÈSE SIMPLIFICATRICE : NOUS SUPPOSONS QUE POUR TOUTE VALEUR DE x LES ε SONT **INDÉPENDANTS ET NORMAUX**, QU'ILS ONT **LE MÊME ÉCART-TYPE** $\sigma = \sigma(\varepsilon)$ **ET UNE MOYENNE** $\mu = 0$.





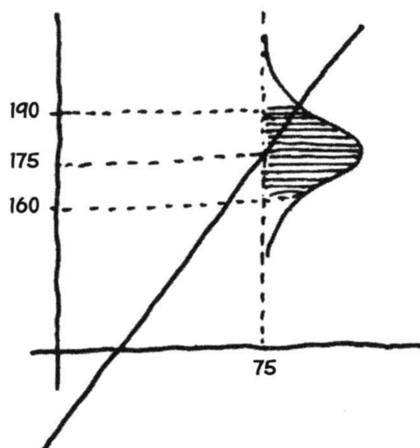
PAR EXEMPLE, SI LE VRAI MODÈLE DE POIDS EST LE SUIVANT :

$$Y = -125 + 4x + \varepsilon$$

OÙ ε SUIT UNE LOI NORMALE AVEC $\mu = 0$ ET $\sigma = 15$ LIVRES, ALORS, SELON CE MODÈLE, LES ÉTUDIANTS MESURANT 75 POUCES (1,90 m) ONT UNE DISTRIBUTION DE :

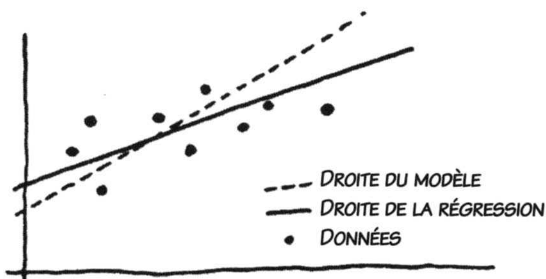
$$Y = -125 + 4 \times 75 + \varepsilon$$
$$= 175 + \varepsilon$$

AINSI POUR $x = 75$, Y SUIT UNE LOI NORMALE DE MOYENNE 175 LIVRES ET D'ÉCART-TYPE 15 LIVRES.



MAINTENANT, ÉTANT DONNÉ NOTRE MODÈLE $Y = \alpha + \beta x + \varepsilon$, NOUS VOULONS, COMME NOUS L'AVONS FAIT DANS LES PRÉCÉDENTS CHAPITRES, PRENDRE UN **ÉCHANTILLON** ET L'UTILISER POUR **ESTIMER** α ET β.

ON PEUT MONTRER QUE α ET β, LES COEFFICIENTS DE LA DROITE DES MOINDRES CARRÉS $y = a + bx$ SONT **BLUE : B**EST **L**INEAR **U**NBIASED **E**STIMATORS, DONC LES MEILLEURS ESTIMATEURS LINÉAIRES NON BIAISÉS DE α ET β.



GARANTIE INCONDITIONNELLE !



COMME D'HABITUDE, DES ÉCHANTILLONS DIFFÉRENTS DONNANT LIEU À DES OBSERVATIONS DIFFÉRENTES NE GÉNÉRERAIENT PAS LA MÊME DROITE DE RÉGRESSION. CES DROITES SONT **DISTRIBUÉES** AUTOUR DE LA DROITE $Y = \alpha + \beta x + \varepsilon$. NOTRE QUESTION EST : COMMENT a ET b SONT-ILS DISTRIBUÉS AUTOUR DE α ET β RESPECTIVEMENT, ET COMMENT CONSTRUIRE DES **INTERVALLES DE CONFIANCE** ET DES **TESTS D'HYPOTHÈSES** ?

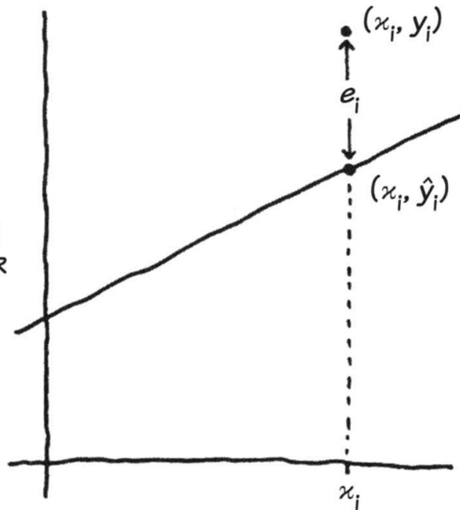


POUR CHAQUE OBSERVATION (x_i, y_i), NOUS AVONS :

$$y_i = a + bx_i + e_i$$

OÙ $e_i = y_i - \hat{y}_i$ EST L'ÉCART ENTRE y_i ET LA DROITE DE RÉGRESSION. LES e_i SONT DES **VALEURS D'ÉCHANTILLON** DE ε ET ELLES DONNENT UN ESTIMATEUR s POUR $\sigma(\varepsilon)$.

$$s = \sqrt{\frac{\sum e_i^2}{n-2}}$$



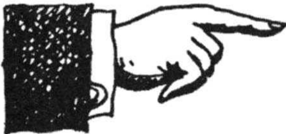
(ON MET $n - 2$ AU DÉNOMINATEUR CAR ON UTILISE DEUX DEGRÉS DE LIBERTÉ POUR LE CALCUL DE a ET b, LAISSANT $n - 2$ INFORMATIONS INDÉPENDANTES POUR ESTIMER σ.)

BIEN QUE CELA PUISSE PARAÎTRE MOINS ÉVIDENT, ON PEUT AUSSI ÉCRIRE s COMME CECI :

$$s = \sqrt{\frac{sc_{yy} - bsc_{xy}}{n-2}}$$

CETTE FORMULE NOUS PERMET DE CALCULER s DIRECTEMENT À PARTIR DES STATISTIQUES D'ÉCHANTILLON.





RÉPÉTONS : s EST UN ESTIMATEUR DU DEGRÉ D'ÉPARPILLEMENT DES DONNÉES AUTOUR DE LA DROITE.

INTERVALLES DE CONFIANCE

LES INTERVALLES DE CONFIANCE À 95 % POUR α ET β ONT LA FORME FAMILIÈRE (VOIR P. 133) SUIVANTE :

$$\beta \in b \pm t_{0,025}s(b)$$

$$\alpha \in a \pm t_{0,025}s(a)$$

OÙ ON UTILISE UN t DE STUDENT AVEC $n - 2$ DEGRÉS DE LIBERTÉ POUR LA MÊME RAISON QUE PRÉCÉDEMMENT.



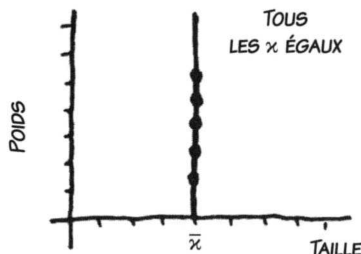
LES ERREURS-TYPES, PAR CONTRE, NE SONT PAS COMME D'HABITUDE. VOICI LES FORMULES (SANS DÉMONSTRATIONS) :

$$s(b) = \frac{s}{\sqrt{sc_{xx}}}$$

$$s(a) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{sc_{xx}}}$$



QU'EST-IL ARRIVÉ À NOTRE PRÉCIEUX $1 / \sqrt{n}$? IL A ÉTÉ REMPLACÉ PAR sc_{xx}. COMME n, sc_{xx} AUGMENTE AVEC LE NOMBRE DE DONNÉES, MAIS IL REFLÈTE AUSSI LA **VARIATION TOTALE** DES DONNÉES EN x. PAR EXEMPLE, SI TOUS LES ÉTUDIANTS DE L'ÉCHANTILLON AVAIENT LA **MÊME TAILLE**, NOUS NE POURRIONS EN TIRER AUCUNE CONCLUSION SUR LA DÉPENDANCE DU POIDS PAR RAPPORT À LA TAILLE. DANS CE CAS, $sc_{xx} = 0$, CE QUI IMPLIQUERAIT $s(b) = \infty$ ET DES INTERVALLES DE CONFIANCE DE **LARGEUR INFINIE**.





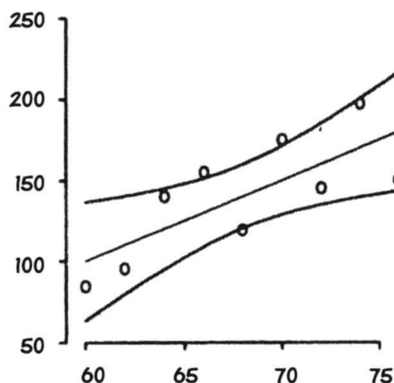
PLUS DE QUESTIONS :

COMMENT PRÉDIRE LA **MOYENNE DES RÉPONSES** Y POUR UNE VALEUR FIXÉE DE x_0 ? PAR EXEMPLE, QUEL EST LE POIDS MOYEN DES ÉTUDIANTS DE 76 POUCES ? L'INTERVALLE DE CONFIANCE À 95 % EST :

$$\alpha + \beta x_0 \in a + bx_0 \pm t_{0,025} s(\hat{y})$$

OÙ :

$$s(\hat{y}) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{sc_{xx}}}$$



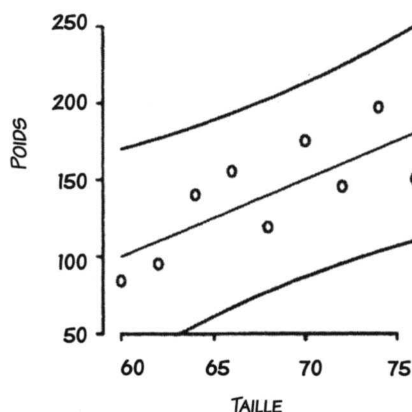
SUPPOSONS QU'UN NOUVEL ÉTUDIANT DE TAILLE x_{nv} ARRIVE. QUELLE PRÉVISION POUVONS-NOUS FAIRE SUR SON POIDS y_{nv} ?

L'INTERVALLE DE PRÉDICTION À 95 % DE y_{nv} POUR L'INDIVIDU DE TAILLE x_{nv} EST :

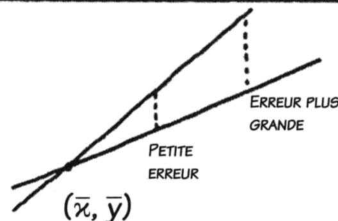
$$y_{nv} \in a + bx_{nv} \pm t_{0,025} s(y_{nv})$$

OÙ

$$s(y_{nv}) = s\sqrt{1 + \frac{1}{n} + \frac{(x_{nv} - \bar{x})^2}{sc_{xx}}}$$



CES DEUX ERREURS-TYPES CONTIENNENT UN TERME CROISSANT EN x (RESPECTIVEMENT EN x_0 ET x_{nv}) QUAND ON S'ÉLOIGNE DE LA MOYENNE $\bar{x}$. POURQUOI L'ERREUR AUGMENTE-T-ELLE EN S'ÉLOIGNANT DE $\bar{x}$? PARCE QUE COMME LA DROITE PASSE TOUJOURS PAR LE POINT $(\bar{x}, \bar{y})$, SI L'ON FAIT PIVOTER CETTE DROITE DE RÉGRESSION ALORS L'IMPACT EST PLUS ÉLEVÉ SUR LES VALEURS ÉLOIGNÉES DE LA MOYENNE !



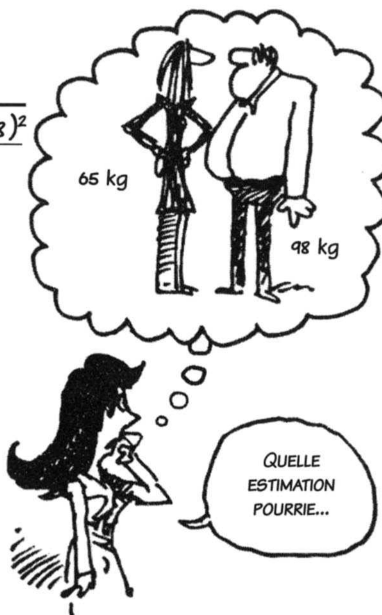
REGARDONS CELA AVEC NOS DONNÉES ARRANGÉES. POUR LA **MOYENNE DES POIDS** POUR UNE TAILLE $x = 76$ POUCES, NOUS AVONS $b = -200$ ET $a = 5$. DONC :

$$Y \in -200 + 5 \times 76 \pm 2{,}365 \times 25{,}15 \sqrt{\frac{1}{9} + \frac{(76-68)^2}{240}}$$

$$Y \in 180 \pm 2{,}365 \times 25{,}15 \sqrt{0{,}3777}$$

$$Y \in 180 \pm 36{,}55 \text{ LIVRES}$$

$$Y \in 81{,}6 \pm 16{,}6 \text{ kg}$$

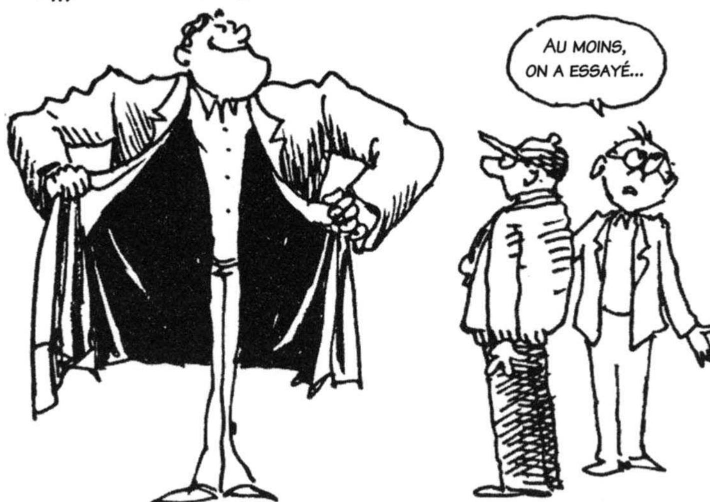


LA MOYENNE ESTIMÉE DU POIDS DES ÉTUDIANTS DE 1,93 m EST DE 81,6 kg, ET NOUS SOMMES SÛRS À 95 % QU'ELLE NE DÉVIERA PAS DE PLUS DE **16,6 kg.**

SI L'ON PREND UN NOUVEL ÉTUDIANT DE 76 POUCES, NOUS UTILISONS NOS DONNÉES ARRANGÉES DE NEUF OBSERVATIONS POUR PRÉDIRE QUE

$$y_{nv} \in -200 + 5 \times 76 \pm 2{,}365 \times 25{,}15 \sqrt{1 + \frac{1}{9} + \frac{(76-68)^2}{240}}$$

$$y_{nv} \in 180 \pm 2{,}365 \times 25{,}15 \times 1{,}174$$
$$y_{nv} \in 180 \pm 70 \text{ LIVRES}$$
$$y_{nv} \in 81{,}6 \pm 31{,}7 \text{ kg}$$

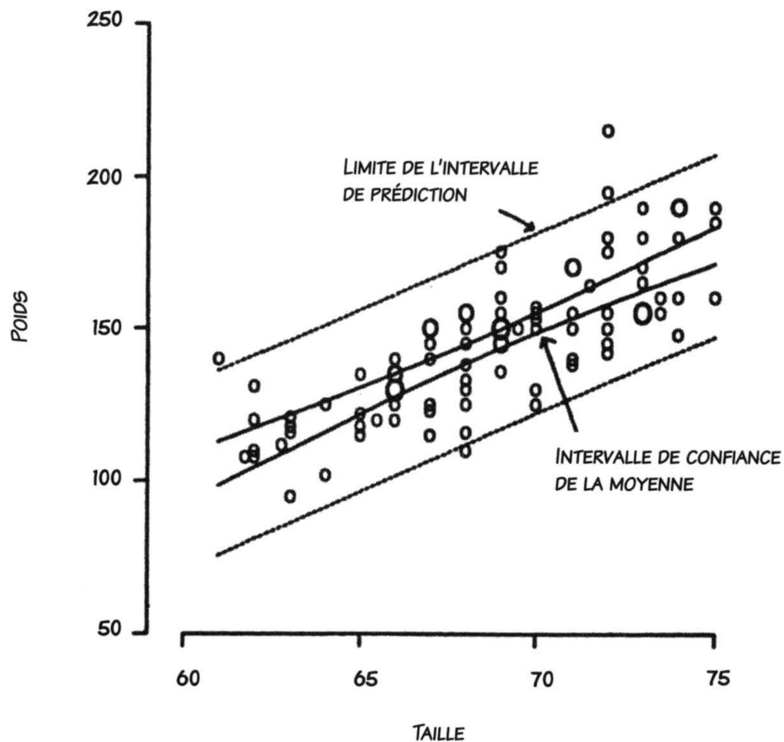


ON PEUT DIRE À L'ENTRAÎNEUR DE FOOT QUE NOUS SOMMES PRESQUE SÛRS QUE CE NOUVEL ÉTUDIANT PÈSE ENTRE **50** ET **113 kg** !

LES INTERVALLES NE SONT PAS FAMEUX ! QUEL EST LE PROBLÈME ?
EN FAIT, IL Y A DEUX PROBLÈMES.

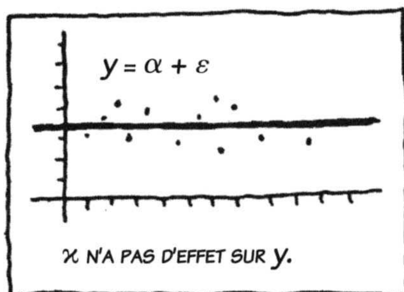


L'ENSEMBLE DES ÉTUDIANTS DE PENN STATE DONNE DE MEILLEURS RÉSULTATS.



Tests d'hypothèses

QUELQU'UN DE COMPLÈTEMENT SCEPTIQUE POURRAIT SUGGÉRER QU'IL N'Y A **PAS DE RELATION** ENTRE LA TAILLE ET LE POIDS. CE QUI REVIENT À DIRE QUE $\beta = 0$.





NOUS PRENONS CECI COMME NOTRE **HYPOTHÈSE NULLE.**

$$H_0 : \beta = 0$$

DANS CE CAS, LA STATISTIQUE DE TEST EST :

$$t = \frac{b}{s(b)}$$

C'EST UN t DE STUDENT AVEC $n - 2$ DEGRÉS DE LIBERTÉ. COMME D'HABITUDE, LA SIGNIFICATION DU TEST DÉPEND DE L'HYPOTHÈSE ALTERNATIVE :

$t > t_\alpha$ POUR $H_a : \beta > 0$

$t < t_\alpha$ POUR $H_a : \beta < 0$

$|t| > |t_{\alpha/2}|$ POUR $H_a : \beta \neq 0$

POUR LES DONNÉES ARRANGÉES DE POIDS, NOUS SUSPECTONS FORTEMENT QUE L'HYPOTHÈSE ALTERNATIVE DOIT ÊTRE :

$$H_a : \beta > 0$$

ON TESTE ALORS :

$$t_{obs} = \frac{5}{s(b)} = \frac{5}{1{,}62}$$

$$\mathbf{= 3{,}08}$$

AVEC 7 DEGRÉS DE LIBERTÉ, $t_{0,05} = 1{,}895$. COMME $t_{obs} > t_{0,05}$, ON REJETTE L'HYPOTHÈSE NULLE AU SEUIL DE SIGNIFICATION DE 5 %, ET ON CONCLUT QU'IL Y A UNE RELATION POSITIVE SIGNIFICATIVE ENTRE LA TAILLE ET LE POIDS.



Régression linéaire multiple

LE MÊME TYPE D'IDÉES BASIQUES PEUT ÊTRE UTILISÉ POUR ANALYSER LES RELATIONS ENTRE UNE VARIABLE DÉPENDANTE ET **PLUSIEURS** VARIABLES INDÉPENDANTES.

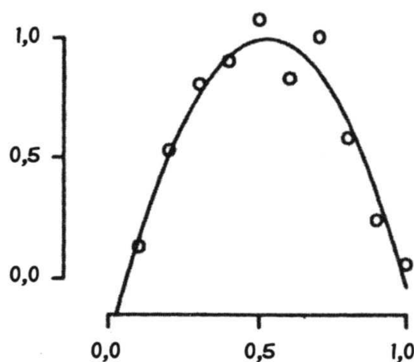
$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

PAR EXEMPLE, LE POIDS EST DÉTERMINÉ PAR UN NOMBRE DE FACTEURS AUTRES QUE LA TAILLE COMME L'ÂGE, LE SEXE, LES RÉGIMES, LE TYPE PHYSIQUE, ETC.



L'ALGÈBRE MATRICIELLE ET LES ORDINATEURS SE COMBINENT POUR FACILITER L'ANALYSE DE CES PROBLÈMES.

Régression non linéaire



PARFOIS LES DONNÉES SUIVENT DE FAÇON ÉVIDENTE UNE COURBE **NON LINÉAIRE.** LES STATISTICIENS ONT DES TAS D'ASTUCES POUR UTILISER DES TECHNIQUES DE RÉGRESSIONS **LINÉAIRES** POUR DES PROBLÈMES **NON LINÉAIRES.** LA PLUS SIMPLE EST DE CONSIDÉRER Y COMME UN POLYNÔME :

$$Y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

ET DE TRAITER x ET x^2 COMME DES VARIABLES INDÉPENDANTES D'UN MODÈLE LINÉAIRE.

Diagnostic des régressions

AJUSTER UN MODÈLE COMPLEXE À DES DONNÉES PEUT PARFOIS OCCULTER DES DIFFICULTÉS OU DES PROBLÈMES. ON UTILISE DES PROCÉDURES DE DIAGNOSTIC DE RÉGRESSION POUR RÉVÉLER TOUTES LES SURPRISES CACHÉES DÉSAGRÉABLES.

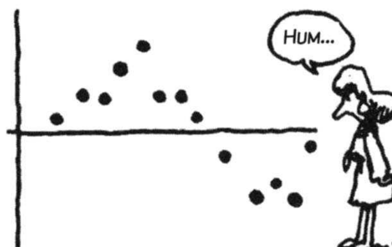


LA MÉTHODE LA PLUS SIMPLE EST DE DESSINER LES ERREURS OU **RÉSIDUS e_i** EN FONCTION DE LA VARIABLE **EXPLICATIVE y_i**. RAPPELEZ-VOUS QUE L'ERREUR ε ÉTAIT CENSÉE ÊTRE INDÉPENDANTE DE x.

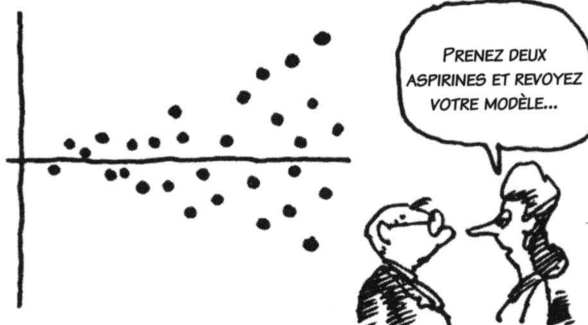
UN NUAGE **ALÉATOIRE** DE POINTS INDIQUE QUE LES HYPOTHÈSES DU MODÈLE SONT SÛREMENT CORRECTES.



TOUT AUTRE **SCHÉMA** INDIQUE UN VÉRITABLE PROBLÈME AVEC LES HYPOTHÈSES DU MODÈLE.



UNE SURPRISE DÉSAGRÉABLE CLASSIQUE (QUE L'ON RETROUVE DANS NOS DONNÉES DE TAILLE/POIDS) EST QUE LES ERREURS SOIENT **HÉTÉROSCÉDASTIQUES,** QUAND LA VARIATION DE e CROÎT LORSQUE y AUGMENTE.



DANS CE CHAPITRE, NOUS AVONS RÉSUMÉ LES IDÉES DE BASE ET LES TECHNIQUES DE L'ANALYSE DE RÉGRESSION, QUI ÉTUDIE LES RELATIONS STATISTIQUES ENTRE DES VARIABLES. AINSI S'ACHÈVE NOTRE DISCUSSION DÉTAILLÉE SUR LES MÉTHODES STATISTIQUES DE BASE. DANS LE DERNIER CHAPITRE, NOUS EXAMINERONS BRIÈVEMENT QUELQUES AUTRES SUJETS ET PROBLÈMES.

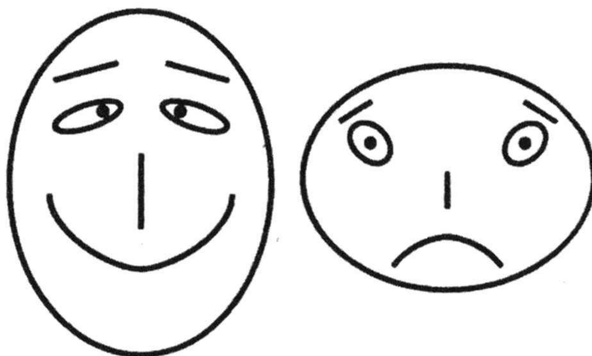


Chapitre 12
CONCLUSION



AFFICHAGE DE DONNÉES

x = INCLINAISON DES SOURCILS
y = TAILLE DES YEUX
z = TAILLE DU NEZ
t = TAILLE DE LA BOUCHE
β = HAUTEUR DU VISAGE
ETC.



ON A VU COMMENT REPRÉSENTER UNE VARIABLE AVEC UN GRAPHE DE POINTS ET DEUX VARIABLES AVEC UN NUAGE DE POINTS. MAIS COMMENT REPRÉSENTER GRAPHIQUEMENT **PLUS DE DEUX** VARIABLES SUR UNE PAGE PLANE ? PARMI LES POSSIBILITÉS, UN GUIDE EN BD SE DOIT DE MENTIONNER L'IDÉE SIMPLE DE **HERMAN CHERNOFF** (1923-) : UTILISER UN VISAGE HUMAIN OÙ CHAQUE TRAIT CORRESPOND À UNE VARIABLE POUR CRÉER UN **VISAGE DE CHERNOFF**.

Analyse statistique de
DONNÉES MULTIVARIÉES

IL EXISTE DES MODÈLES MULTIVARIÉS POUR AIDER À L'ANALYSE ET À L'AFFICHAGE DE DONNÉES À n DIMENSIONS. VOICI QUELQUES TECHNIQUES MULTIVARIÉES :

L'analyse typologique

VISE À DIVISER LA POPULATION EN SOUS-GROUPES HOMOGÈNES. PAR EXEMPLE, QUAND ON ANALYSE LES SCHÉMAS DE VOTE DU CONGRÈS, ON PEUT VOIR QUE LES REPRÉSENTANTS DU SUD-EST ET CEUX DU NORD-EST FORMENT DEUX GROUPES DISTINCTS.



L'analyse discriminante

EST LE PROCÉDÉ INVERSE. PAR EXEMPLE, LE BUREAU D'ADMISSION EN LICENCE AIMERAIT DISPOSER DE DONNÉES QUI LUI INDIQUERAIENT PAR AVANCE SI LE CANDIDAT **RÉUSSIRA** SON DIPLÔME (CONTRIBUANT ALORS AU FONDS DES ANCIENS) OU S'IL SERA EN **ÉCHEC** (QUITTANT L'UNIVERSITÉ POUR UNE RECONVERSION DANS L'HUMANITAIRE).



L'analyse factorielle

VISE À EXPLIQUER DES DONNÉES À DIMENSIONS MULTIPLES AVEC UN PETIT NOMBRE DE VARIABLES. UN PSYCHOLOGUE PEUT PAR EXEMPLE PROPOSER UN TEST DE 100 QUESTIONS EN PENSANT QUE LES RÉPONSES NE DÉPENDENT QUE DE QUELQUES FACTEURS COMME L'EXTRAVERSION, L'AUTORITARISME ET L'ALTRUISME. ON PEUT ALORS RÉSUMER LES RÉSULTATS DU TEST À PARTIR DE SCORES COMBINÉS DANS CES DIMENSIONS.

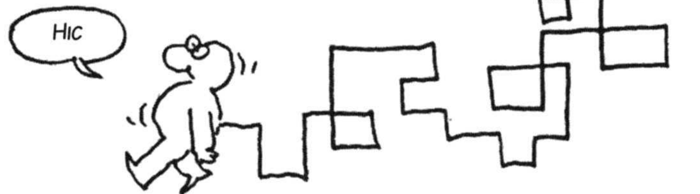


IL Y A AUSSI PLUS À FAIRE AVEC LES

PROBABILITÉS :

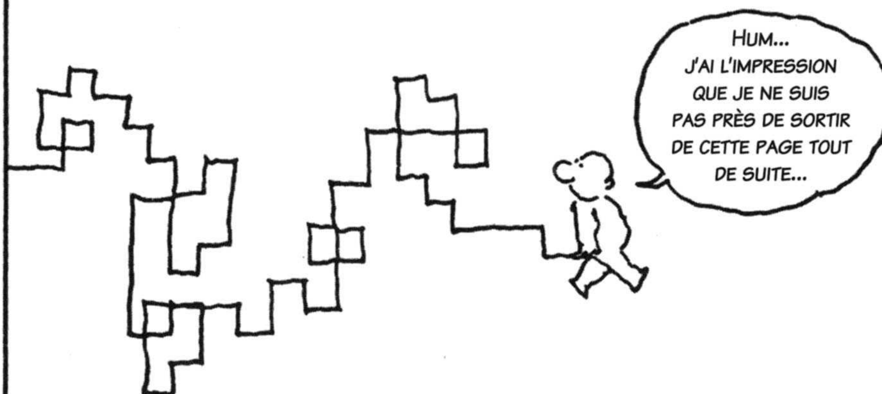
Une marche aléatoire

ELLE COMMENCE PAR UN LANCER DE PIÈCE. SUPPOSONS QUE VOUS AVANCIEZ SI C'EST FACE ET QUE VOUS RECULIEZ SI C'EST PILE (AVEC DEUX PIÈCES ON PEUT FAIRE CELA SUR DEUX DIMENSIONS). CES LANCERS RÉPÉTÉS PRODUISENT UN PROCESSUS STOCHASTIQUE APPELÉ MARCHE ALÉATOIRE. LES MODÈLES DE MARCHE ALÉATOIRE SONT UTILISÉS POUR LES NÉGOCIATIONS EN BOURSE ET LA GESTION DE PORTEFEUILLE.



L'analyse de séries temporelles

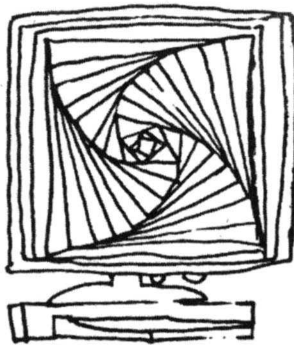
ELLE TRAITE D'ENSEMBLES DE DONNÉES, COMME LES MARCHES ALÉATOIRES, QUI VARIENT AU COURS DU TEMPS : LES TEMPÉRATURES GLOBALES ET LOCALES, LE PRIX DU PÉTROLE, ETC. L'ANALYSE DE SÉRIES TEMPORELLES SE FAIT À L'AIDE DE MODÈLES ALÉATOIRES AFIN DE PRÉDIRE LES VALEURS FUTURES.



NOUS AVONS DÉJÀ VU L'AIDE APPORTÉE PAR LES ORDINATEURS POUR L'ARITHMÉTIQUE ET L'ANALYSE. IL Y A AUSSI DES IDÉES STATISTIQUES DONT L'**EXISTENCE** MÊME PROVIENT DES ORDINATEURS.

Analyse d'image

UNE IMAGE D'ORDINATEUR EST CONSTITUÉE DE PLUSIEURS MILLIONS DE PIXELS (ÉLÉMENTS DE L'IMAGE). UNE IMAGE DONNÉE EST DONC UN ENSEMBLE HAUTEMENT STRUCTURÉ DE MILLIONS DE VECTEURS DE PIXELS. L'ANALYSE D'IMAGES TEND À EXTRAIRE DU SENS DE CE TYPE D'INFORMATION.





Ré-échantillonnage

PARFOIS, LES ÉCARTS-TYPES ET LES BORNES DE CONFIANCE SONT IMPOSSIBLES À DÉTERMINER. ON UTILISE ALORS LE RÉ-ÉCHANTILLONNAGE, UNE TECHNIQUE QUI SE SERT DE L'ÉCHANTILLON LUI-MÊME **COMME S'IL S'AGISSAIT D'UNE POPULATION.** LES TECHNIQUES S'APPELLENT **RANDOMISATION, JACKKNIFE,** ET TECHNIQUE DU **BOOTSTRAPPING.**



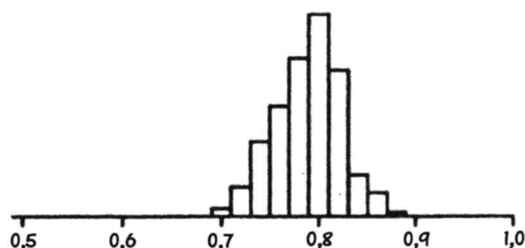
Ré-échantillonnage (suite)

POUR RÉ-ÉCHANTILLONNER, L'ORDINATEUR

- RE-ÉCHANTILLONNE L'ÉCHANTILLON ;
- CALCULE LES ESTIMATIONS DE CES NOUVEAUX ÉCHANTILLONS ;
- RÉPÈTE PLUSIEURS FOIS LES DEUX PREMIÈRES ÉTAPES, ET ESTIME LA VARIANCE DES ESTIMATIONS DES ÉCHANTILLONS DE L'ÉCHANTILLON.



VOUS VOUS SOUVENEZ DU COEFFICIENT DE CORRÉLATION r DES 92 POIDS ET TAILLES D'ÉTUDIANTS DU CHAPITRE 11 ? QUEL EST L'**ÉCART-TYPE** DE r ? L'ORDINATEUR PREND 200 ÉCHANTILLONS (« BOOTSTRAP ») DE CES 92 OBSERVATIONS, CALCULE r CHAQUE FOIS ET TRACE UN HISTOGRAMME DES r TROUVÉS.





CORRÉLATION DE BOOTSTRAPS

REMARQUEZ QUE LA VARIANCE DES ESTIMATIONS DES BOOTSTRAPS EST RELATIVEMENT FAIBLE.



LA QUALITÉ DES DONNÉES



VRAISEMBLABLEMENT DES PETITES ERREURS DANS L'ÉCHANTILLONNAGE, DANS LES MESURES ET L'ENREGISTREMENT DES DONNÉES PEUVENT CAUSER DES RAVAGES SUR N'IMPORTE QUELLE ANALYSE. **R. A. FISHER** (1890-1962), GÉNÉTICIEN ET FONDATEUR DES STATISTIQUES MODERNES, NE SE CONTENTAIT PAS DE CONCEVOIR ET D'ANALYSER DES EXPÉRIENCES CONCERNANT L'ÉLEVAGE DES ANIMAUX. IL NETTOYAIT AUSSI LEURS CAGES ET LES SOIGNAIT CAR IL SAVAIT QUE LA PERTE DE L'UN D'ENTRE EUX RISQUAIT D'AFFECTER SES RÉSULTATS.

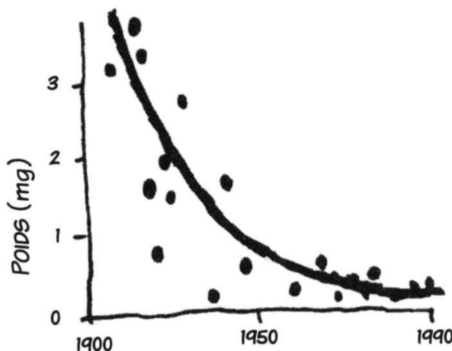
AVEC LEURS ORDINATEURS, LEURS BASES DE DONNÉES ET LEURS SUBVENTIONS GOUVERNEMENTALES, LES STATISTICIENS MODERNES ONT PERDU L'HABITUDE DE METTRE LA MAIN À LA PÂTE.





SI VOUS TRACIEZ LA MASSE DE RÉSIDUS DE RATS SOUS LES ONGLES DES STATISTICIENS AU COURS DU SIÈCLE DERNIER, VOUS AURIEZ SÛREMENT UN GRAPHIQUE DE CE TYPE :





Innovation

LES MEILLEURES SOLUTIONS NE SONT PAS TOUJOURS DANS LES LIVRES ! PAR EXEMPLE, UNE COMPAGNIE RECRUTÉE POUR ESTIMER LA COMPOSITION D'UNE **DÉCHARGE PUBLIQUE** A ÉTÉ CONFRONTÉE À DES PROBLÈMES INTÉRESSANTS MAIS QUE VOUS NE TROUVEREZ PAS DANS LES TEXTES STANDARDS...



Communication

LES ANALYSES BRILLANTES SONT INUTILES SI LES RÉSULTATS NE SONT PAS CLAIREMENT EXPRIMÉS AVEC DES MOTS SIMPLES, Y COMPRIS SUR LE DEGRÉ D'INCERTITUDE DES CONCLUSIONS. PAR EXEMPLE, LES MÉDIAS PARLENT MAINTENANT DAVANTAGE DES MARGES D'ERREUR DE LEURS SONDAGES.



Travail d'équipe

DANS NOTRE SOCIÉTÉ COMPLEXE, LA RÉSOLUTION DE NOMBREUX PROBLÈMES REQUIERT UN TRAVAIL D'ÉQUIPE. LES INGÉNIEURS, LES STATISTICIENS ET LES OUVRIERS À LA CHAÎNE COOPÈRENT POUR ACCROÎTRE LA QUALITÉ DE LEURS PRODUITS. LES BIOSTATISTICIENS, LES MÉDECINS ET LES MILITANTS ONT TRAVAILLÉ ENSEMBLE POUR CONCEVOIR DES ESSAIS CLINIQUES AFIN D'ÉVALUER RAPIDEMENT L'EFFICACITÉ DE MÉDICAMENTS DESTINÉS À COMBATTRE LE SIDA.



EH BIEN VOILÀ ! MAINTENANT, VOUS DEVRIEZ POUVOIR FAIRE À PEU PRÈS N'IMPORTE QUOI AVEC LES STATISTIQUES, EXCEPTÉ **MENTIR, TRICHER, VOLER** ET PARIER OU JOUER.




AVEZ-VOUS UNE ASSURANCE CONVENABLE POUR FAUTE PROFESSIONNELLE STATISTIQUE ?

BIBLIOGRAPHIE



POUR BIEN DÉBUTER

HAHN C., MACÉ S., ***MÉTHODES STATISTIQUES APPLIQUÉES AU MANAGEMENT***, PARIS, PEARSON, 2012. PRÉSENTATION DES MÉTHODES STATISTIQUES SANS TROP DE FORMALISME ET PLUTÔT AXÉE SUR DES EXEMPLES PRATIQUES.

VESSEREAU A., ***LA STATISTIQUE***, PARIS, PUF, COLL. « QUE SAIS-JE ? », 2002. UN PEU DE STATISTIQUES DANS UN FORMAT DE POCHE.

POUR APPROFONDIR ET POUR L'ÉTUDIANT

ANDERSON D. R., CAMM J. D., COCHRAN J. J. ET SWEENEY. D. J., ***STATISTIQUES POUR L'ÉCONOMIE ET LA GESTION***, 5e ÉDITION, DE BOECK, 2015. UN MANUEL TRÈS COMPLET AVEC DE NOMBREUX EXEMPLES ET EXERCICES.

LECOUTRE J.-P., ***STATISTIQUE ET PROBABILITÉS***, 2e ÉDITION, PARIS, DUNOD, 2003. COURS AVEC EXERCICES CORRIGÉS D'UN NIVEAU DE DEUXIÈME ANNÉE DE LICENCE D'ÉCONOMIE-GESTION.

PHAN T., ROWENCZYK J.-P., ***EXERCICES ET PROBLÈMES DE STATISTIQUE ET PROBABILITÉS***, 2e ÉDITION, PARIS, DUNOD, 2007. UN AUTRE COURS AVEC EXERCICES ET PROBLÈMES CORRIGÉS S'ADRESSANT À UN PUBLIC PLUS SCIENTIFIQUE QUE LE PRÉCÉDENT.



EN ANGLAIS

POUR L'ÉTUDIANT

FREEDMAN D., PISANI R. ET PURVES R., ***STATISTICS***, 4th EDITION, NEW YORK, W. W. NORTON, 2007. INTRODUCTION APPROFONDIE AUX STATISTIQUES UTILISANT LE MOINS POSSIBLE LES MATHÉMATIQUES ET LES ÉQUATIONS.

MOORE D. S., NOTZ W. I., ***STATISTICS CONCEPTS AND CONTROVERSIES***, NEW YORK, W. H. FREEMAN, 2008. SOULIGNE LES IDÉES PLUTÔT QUE LE CÔTÉ MÉCANIQUE.

CES DEUX LIVRES SONT DES CLASSIQUES : BONS, LITTÉRAIRES ET SPIRITUELS. À CÔTÉ DE CEUX-CI, IL EXISTE DES CENTAINES DE LIVRES DISPONIBLES ET LA PLUPART D'ENTRE EUX SONT, SELON NOUS, SATISFAISANTS.



COMMENT MENTIR, TRICHER ET PARIER :

VOS AUTEURS, QUI SONT DES SAINTS, MANQUENT D'EXPÉRIENCE EN LA MATIÈRE. VOICI QUELQUES CONSEILS DE PROS :

HUFF D., ***HOW TO LIE WITH STATISTICS***, NEW YORK, W. W. NORTON, 1954. BON MARCHÉ ET TOUJOURS PUBLIÉ.

ORKIN M., ***CAN YOU WIN ?***, NEW YORK, W. H. FREEMAN, 1991. CONSEILS D'UN EXPERT DES PROBABILITÉS ET DES JEUX (N'EST PLUS PUBLIÉ, MAIS ON TROUVE ENCORE FACILEMENT DES EXEMPLAIRES D'OCCASION).

SPIRER H. F., SPIRER L. ET JAFFE A. J., ***MISUSED STATISTICS***, 2nd EDITION, NEW YORK, MARCEL DECKER, 1998. UN LIVRE FAISANT PARTIE D'UNE COLLECTION RECONNUE SUR LES STATISTIQUES.

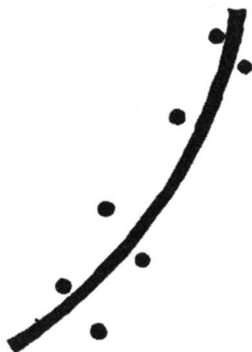
Droit et société



Gastwirth J. L., ***Statistical Reasoning in Law and Policy***, vol. 1 et 2, San Diego, Academic Press, 1988. Les détails pratiques de la loi incluant le cas de sélection de jury décrit au début du chapitre 9.

Dans le chapitre 9, le commentaire non judiciaire sur le poker est tiré d'une affaire réelle ; nous en avons eu la confirmation par le Dr John de Cani de l'université de Pennsylvanie.

Le comité de pilotage du groupe de recherche et d'étude sur la santé des médecins, « **Final report on the Aspirin component of the ongoing physicians' health study** », *The New England Journal of Medicine*, vol. 321, p. 129-135. Vous y trouverez des détails concernant l'étude sur l'aspirine décrite dans le chapitre 9.



Description graphique des données

Tufte E. R., ***The Visual Display of Quantitative Information***, New Haven, Graphics Press, 2001. Les livres de Tufte ont fixé de nouveaux standards pour la communication des données quantitatives.

Cleveland W. S., ***The Elements of Graphing Data***, Summit (N. J.), Hobart Press, 1994. La plupart des idées graphiques décrites ici peuvent être implémentées avec des logiciels de statistiques comme *R* ou d'autres systèmes.

Histoire



Box J. F., ***R. A. Fisher : The Life of a Scientist***, New York, Wiley, 1978. La biographie faite par la fille de l'un des statisticiens les plus influents et controversés du xxe siècle (elle n'est plus publiée et est devenue plutôt rare et chère).

David F. N., ***Games, Gods and Gambling***, Mineola (N. Y.), Dover 2012. Réimpression d'un livre classique sur l'histoire des statistiques.

Salsburg D., ***The Lady Tasting Tea : How Statistics Revolutionized Science in the Twentieth Century***, New York, Henry Holt & Co., 2002. Une histoire populaire des statistiques modernes.

LOGICIEL DE STATISTIQUES

DANS CE LIVRE, NOUS AVONS UTILISÉ LE LOGICIEL STATISTIQUE **MINITAB** (INC. STATE COLLEGE PA). LES DONNÉES DE POIDS ET DE TAILLES DES ÉTUDIANTS DE PENN STATE PROVIENNENT DE LA BASE DE DONNÉES **PULSE** DE CE SYSTÈME. LA VERSION ACTUELLE POUR ÉTUDIANTS DE MINITAB EST **MINITAB EXPRESS**. LES FICHIERS D'AIDE DE MINITAB SONT CLAIRS ET DÉTAILLÉS, IL S'AGIT PRESQUE EN SOI D'UN COURS BASIQUE DE STATISTIQUES. LES GRAPHIQUES INFORMATIQUES DE CETTE BD ONT ÉTÉ GÉNÉRÉS AVEC **S-PLUS**.

R EST UN SYSTÈME SIMILAIRE ET GRATUIT DE STATISTIQUES, TRÈS LARGEMENT UTILISÉ POUR DES ANALYSES GRAPHIQUES OU NUMÉRIQUES DÉTAILLÉES. R ET S ONT, TOUS LES DEUX, ÉTÉ DÉVELOPPÉS PAR DES STATISTICIENS TRAVAILLANT À AT&T BELL LABS.



R EST MAINTENANT SOUTENU PAR LA COMMUNAUTÉ ACADÉMIQUE EN SCIENCES INFORMATIQUES ET STATISTIQUES PARTOUT DANS LE MONDE. MINITAB EXPRESS ET R FONCTIONNENT TOUS DEUX SOUS PC OU MAC.



VOS PRÉFÉRENCES INDIVIDUELLES OU VOS GOÛTS VOUS AMÈNERONT À CHOISIR L'UN OU L'AUTRE DE CES LOGICIELS. POUR CERTAINS, RIEN NE PEUT BATTRE LE « GRATUIT ET SOPHISTIQUÉ » (C'EST R), POUR D'AUTRES, IL FAUT DU « SIMPLE, DU CLAIR ET DU TRANSPARENT » (C'EST MINITAB).

BIEN D'AUTRES LOGICIELS DE STATISTIQUES EXISTENT. LA PLUPART ONT MIGRÉ OU VONT MIGRER VERS UN CALCUL DANS LE « CLOUD », DE FAÇON À ÉLIMINER LES PROBLÈMES TECHNIQUES ET CEUX DE COMPATIBILITÉ POUR LES ÉTUDIANTS COMME POUR LES ANALYSTES PROFESSIONNELS DE DONNÉES.

LES GRANDS SYSTÈMES DE BIG DATA DE QUALITÉ INDUSTRIELLE COMME **SAS** OU **JUMP** DE SAS INC., **SPSS** D'IBM, ET **STATA** SONT LARGEMENT UTILISÉS DANS LES ENTREPRISES ET LA RECHERCHE. ON LES TROUVE TOUS À DES PRIX RÉDUITS POUR LES ÉTUDIANTS. UN ÉTUDIANT AVISÉ DEVRAIT EN PROFITER POUR APPRENDRE L'UN DE CES SYSTÈMES D'ANALYSE DE QUALITÉ INDUSTRIELLE.

POUR CHAQUE PROGRAMME, IL Y A DE NOMBREUX LIVRES ET MANUELS DÉTAILLÉS. UNE BONNE TRENTAINE DE LIVRES OU DE MANUELS SE PRÉTENDENT « UNE INTRODUCTION À R ». NOTRE LIVRE FAVORI DE PROGRAMMATION STATISTIQUE EST ***THE LITTLE SAS BOOK*** DE L. DELWICHE ET S. SLAUGHTER. ON PEUT Y AJOUTER ***THE MINITAB STUDENT HANDBOOK*** DE B. RYAN, B. JOINER ET J. CRYER.

LES PACKS STATISTIQUES SONT DIFFÉRENTS EN CE QUI CONCERNE DES DÉTAILS PARFOIS IMPORTANTS. VOUS DEVEZ ÊTRE UN ACHETEUR ÉCLAIRÉ. NOUS VOUS RECOMMANDONS DE CHOISIR UN SYSTÈME QUE VOS COLLÈGUES ONT DÉJÀ TESTÉ. PEU D'ENTRE NOUS SONT TAILLÉS POUR ÊTRE DES PIONNIERS DES LOGICIELS. PENDANT L'APPRENTISSAGE D'UN LOGICIEL, UTILISEZ DES ENSEMBLES PETITS DE DONNÉES FAMILIÈRES. **LA PARTIE LA PLUS COÛTEUSE D'UN LOGICIEL EST LE TEMPS QUE VOUS LUI CONSACREREZ.** LA RÈGLE DE BASE POUR APPRENDRE UN LOGICIEL DE STATISTIQUES EST SIMPLE : **LA CONNAISSANCE PRÉCÈDE LES RÉSULTATS.**

CHERCHER À APPRENDRE SIMULTANÉMENT LA THÉORIE STATISTIQUE ET LA PROGRAMMATION STATISTIQUE EST UN PEU COMME ESSAYER DE MARCHER EN MÂCHANT UN CHEWING-GUM. DIFFÉRENTES APTITUDES ET MÉTHODES DE RÉFLEXION SONT CONCERNÉES EN MÊME TEMPS. PRÉVOYEZ DES TEMPS DIFFÉRENTS POUR APPRENDRE CHAQUE THÈME. ENSUITE COMBINEZ L'ÉTUDE DES DEUX À LA FOIS. DE CETTE FAÇON, VOUS POURREZ DEVENIR UN **MÂCHEUR, MARCHEUR, PROGRAMMEUR STATISTICIEN!**



Index

D

E

F

G

H

I

J-L

M

N

O

P

Q-R

S

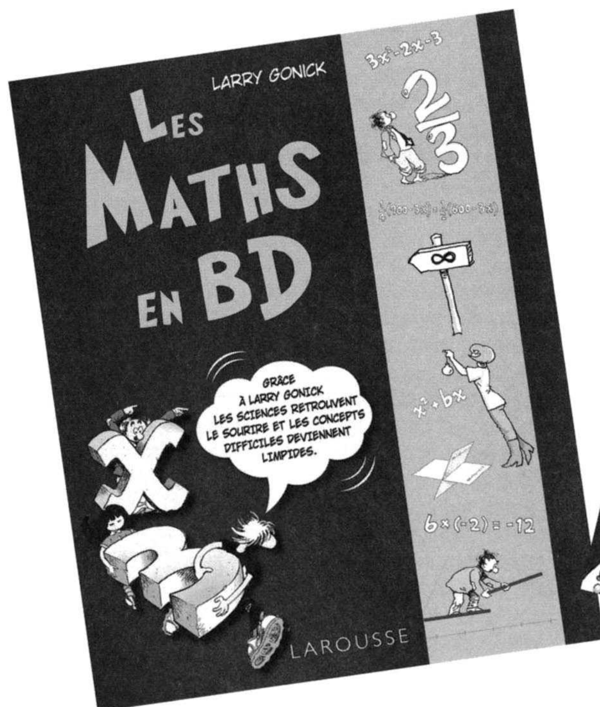
T

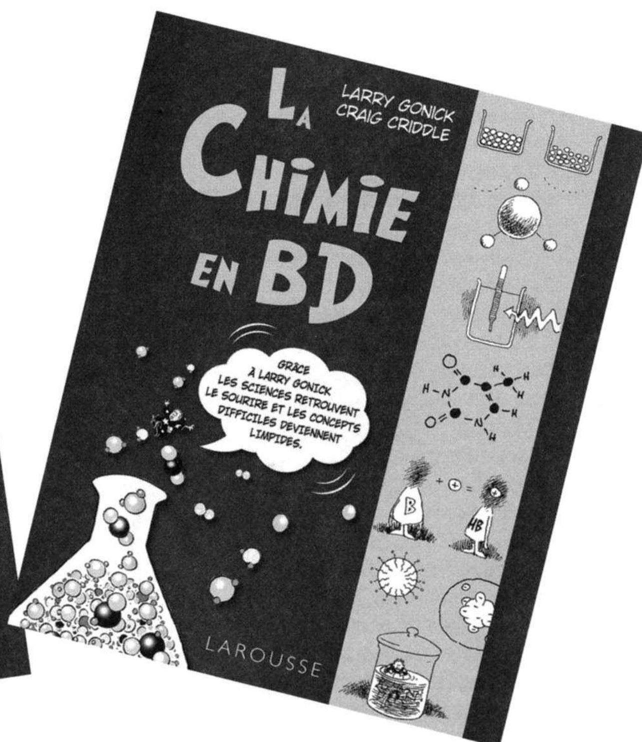
U

V

Z

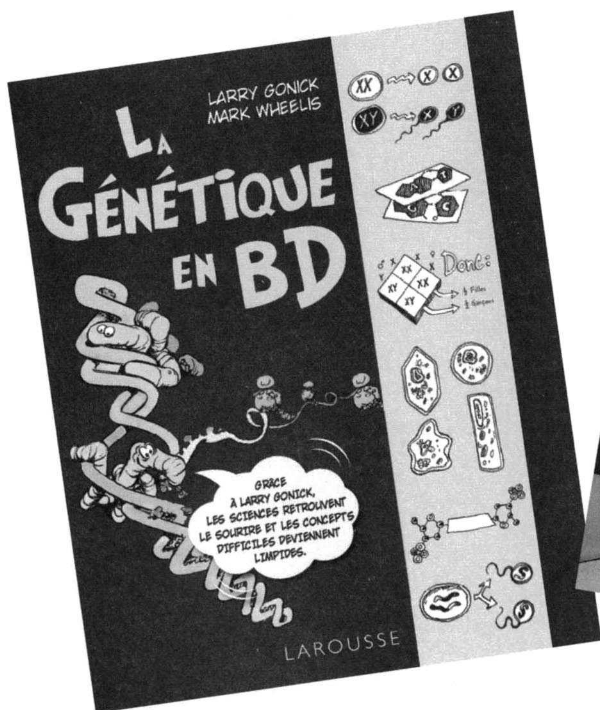
D'AUTRES « SCIENCES EN BD » AVEC LARRY GONICK CHEZ LAROUSSE...

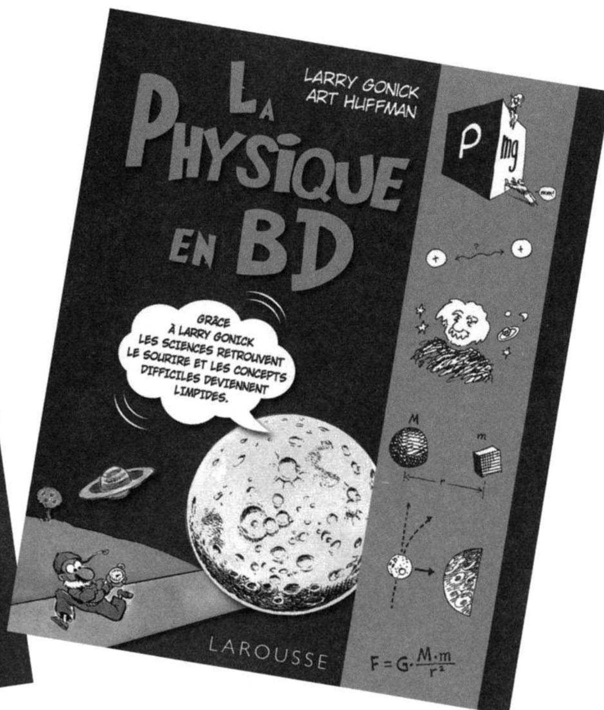




LES MATHS EN BD
LARRY GONICK
AVEC CE VOLUME SPÉCIALEMENT CONSACRÉ À L'ALGÈBRE, RÉVISEZ TOUS LES FONDAMENTAUX : RAPPELS SUR LES NOMBRES ET LES OPÉRATIONS, MISE EN ÉQUATION DES VARIABLES ET SIMPLIFICATIONS, ÉQUATIONS DU PREMIER ET DU SECOND DEGRÉS, GRAPHIQUES, RACINES CARRÉES... AVEC DES CAS CONCRETS ET DES PROBLÈMES CORRIGÉS POUR S'ENTRAÎNER !
240 PAGES

LA CHIMIE EN BD
LARRY GONICK ET CRAIG CRIDDLE
PASSEZ EN REVUE TOUS LES DOMAINES DE LA CHIMIE : LA MATIÈRE ET LES ÉLÉMENTS DU TABLEAU PÉRIODIQUE, LE PROTON, LE NEUTRON ET L'ÉLECTRON (ET SON RÔLE DANS LA LIAISON CHIMIQUE), LES IONS, LES MOLÉCULES, LA MASSE MOLAIRE, LES GAZ PARFAITS, LA THERMODYNAMIQUE... AVEC DES CAS CONCRETS ET DES PROBLÈMES CORRIGÉS POUR S'ENTRAÎNER !
256 PAGES


LARRY GONICK
MARK WHEELIS
LA GÉNÉTIQUE EN BD
GRÂCE À LARRY GONICK, LES SCIENCES RETROUVENT LE SOURIRE ET LES CONCEPTS DIFFICILES DEVIENNENT LIMPIDES.
LAROUSSE


LARRY GONICK
ART HUFFMAN
LA PHYSIQUE EN BD
GRÂCE À LARRY GONICK LES SCIENCES RETROUVENT LE SOURIRE ET LES CONCEPTS DIFFICILES DEVIENNENT LIMPIDES.
LAROUSSE
$F = G\cdot\frac{M\cdot m}{r^2}$



LAROUSSE s'engage pour l'environnement en réduisant l'empreinte carbone de ses livres. Celle de cet exemplaire est de :
1,5 kg éq. CO_2
Rendez-vous sur www.larousse-durable.fr

IMPRIMÉ EN ESPAGNE CHEZ UNIGRAF
DÉPÔT LÉGAL : SEPTEMBRE 2016
318486-01/11033648 – SEPTEMBRE 2016

Les Statistiques en BD

AUJOURD'HUI, PRESQUE TOUS LES DOMAINES D'ACTIVITÉ UTILISENT LES STATISTIQUES : ÉCONOMIE, FINANCE, TECHNOLOGIE, INFORMATIQUE, BIOLOGIE, COMMUNICATION…

AVEC CE VOLUME QUI PASSE EN REVUE LES GRANDS PRINCIPES DE CETTE DISCIPLINE, NOTAMMENT LES PLUS POINTUS, RÉVISEZ VOS FONDAMENTAUX EN VOUS AMUSANT ET RÉVÉLEZ LE PETIT GÉNIE QUI SOMMEILLE EN VOUS.

- *Au programme : données, descriptions, probabilités, écarts-types, variables aléatoires, distributions, lois normale et binomiale, échantillonnages, intervalles de confiance, tests d'hypothèse, méthodes expérimentales, régression linéaire…*
- *Des illustrations qui éclaircissent les énoncés scientifiques tout en y injectant une bonne dose d'humour. Avec ses mises en situation désopilantes et ses cas pratiques, Larry Gonick parvient à rendre les statistiques modernes, utiles et passionnantes.*

LARRY GONICK est un auteur de bandes dessinées et un scientifique américain. Il a enseigné les mathématiques à Harvard et participe au développement du journalisme scientifique au MIT (Massachusetts Institute of Technology). Ses nombreux « comics » de vulgarisation scientifique (chimie, histoire de l'Univers, informatique, physique, génétique…) sont des best-sellers outre-Atlantique, où ils sont devenus des classiques.

WOOLLCOTT SMITH est professeur émérite de statistiques à l'université de Temple à Philadelphie. Il est titulaire d'une licence et d'un master de l'université du Michigan et d'une thèse de l'université Johns-Hopkins.

17,99 € PRIX TTC FRANCE

ISBN 978-2-03-593291-4



Illustrations de couverture : Larry Gonick

LAROUSSE
DICTIONNAIRES

www.editions-larousse.fr